NOTE TO USERS

The original manuscript received by UMI contains pages with indistinct and/or slanted print. Pages were microfilmed as received.

This reproduction is the best copy available

UMI

University of Alberta

Theory of Generalizability and Optimization of Marketing Measurement

by

Ujwal Anilchandra Kayandé

©

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of Doctor of Philosophy

in

Marketing

Faculty of Business

Edmonton, Alberta
Spring 1998

National Library of Canada
Acquisitions and Bibliographic Services
395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale du Canada
Acquisitions et services bibliographiques
395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-29053-0

Canada

University of Alberta

Faculty of Graduate Studies and Research

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled *Theory of Generalizability and Optimization of Marketing Measurement* submitted by Ujwal Anilchandra Kayandé in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Marketing.

Dr. Adam Finn

Dr. Terry Elrod

Dr. Peter Popkowski Leszczyc

Dr. Paul Messinger

Dr. Michael Gillespie

Dr. Roland T. Rust

December 11, 1997

To

Aai-Baba,

who have taught me all that matters

Abstract

The development of methods to acquire information about different characteristics of objects has been an important aspect of scientific inquiry. The focus of this thesis is on the development and study of a class of methods that improve the quantitative measurement of characteristics of marketing objects. Different marketing problems require the measurement and scaling of different characteristics of *different objects*. However, the academic literature on measurement in marketing, through its continued use of classical reliability theory to assess the quality of measurement, has focused on the scaling of only individuals and ignores other objects, even though many purposes of measurement may require the scaling of characteristics of such objects. Generalizability theory is a sophisticated psychometric approach that can be applied to design efficient measurement and explicitly take into account differences in the purpose of measurement. In the first paper of the thesis, a generalizability theory perspective is taken to propose a general framework to optimize the design of marketing measurement for different purposes of measurement. An empirical study on service quality measurement was conducted to illustrate the advantage of the generalizability approach to measurement. The application of the framework requires the extensive use of estimates of variance components. The second paper of the thesis investigates two issues that provide an indication of the methodological limitations of the extant framework. First, does the optimal measurement design depend on the method used to estimate variance components? Second, are the interval estimates such that the *optimality* of a design generated from point estimates might be seriously threatened? To answer these questions, different statistical methods to obtain point and interval estimates of variance components

are reviewed and applied to the service quality data. The third paper of the thesis presents a generalization of the framework to the case of *multidimensional* measurement. Different ways of using multidimensional information are reviewed, the dependence of the reliability assessment method on these different ways of using information is illustrated, methods to assess reliability and design future measurement in each case are presented, and finally the methodological framework is illustrated with the case of measurement of retail store quality.

Acknowledgements

This thesis owes its completion to the efforts of many individuals, some of who are family and some of who are colleagues, and all of who are friends. My academic upbringing owes much to the untiring efforts of Terry Elrod, Richard D. Johnson, Peter Popkowski Leszczyc, Mukesh Bhargava, Dave Jobson, Michael Enzle, and most of all, Adam Finn. All of them have been "spoiling" me for a long time; hopefully, I will have the opportunity to give back to some doctoral student a fraction of what was given to me, without *ever* asking for anything in return. Adam has spent a large proportion of his academic time in the past few years simply indulging my curiosity and my curious ways. Hopefully, the future will show that the time was well spent! In addition, Paul Messinger and Mike Gillespie have been very helpful with the completion of this thesis.

Peter Popkowski Leszczyc has been the kindest of friends, helpful and encouraging when most *and* least needed. The warmth of friends like Anil Sawhney, Julian Andrews, Peter Roberts, Tom Johnson, Carla Carnaghan, Barry Ard, Gerald Haübl, Shauna Crowe, Tony Kurian, and Ben helped me enjoy the cold frigidity of "dissertating" in Edmonton. The PhD Program Office, specially Jeanette Gosine, has been terrific in providing the backbone for me to obtain a complete education.

My family has been most encouraging of my academic ventures, but also tolerant of my procrastinations and other weaknesses. My brother, Udayan, and my sister, Smriti, have been pillars of support for me in my life at and away from home. As siblings, they come no better than those that are mine! My parents, Anil and Shradha Kayande, to whom I have dedicated this thesis, have been the most wonderful parents I could have asked for; they have shown me the way by example, and never by speech. More than

anything, my siblings and parents will be happy to know that "my learning will go on forever . . .". My uncle and aunt, Jivan and Siddhi Kayande, have been very supportive and wonderful through hard and fun times in Edmonton.

The importance weights for the components of a combination are not always optimally determined by eigenvalues and eigenvectors; they are sometimes optimally determined by human nature and emotions. In our perhaps not-so-linear combination, Yoshita is more significant, as is obvious to anyone who knows both of us, because of her affectionate, kind, and loving nature. So finally, and most importantly, I thank my other and more important half, Yoshita, for always being there for me when I most needed a smile and cheer.

Contents

Chapter	Title	Page
1.	Introduction	1
Bibliography		8
2.	Reliability Assessment and Optimization of Marketing Measurement	9
2.1	Using Generalizability Theory to Optimize Measurement	12
2.2	Empirical Illustration with Service Quality Measurement	17
2.2.1	Empirical Investigation	19
2.2.2	Reduction of Error variance in Applied Decision Studies	21
2.2.3	Problem 1: Benchmarking Chains within One Retail Sector	22
2.2.4	Problem 2: Identifying Priorities for Quality Improvement	25
2.2.5	Problem 3: Benchmarking Chains on Different Aspects of Service Quality	26
2.2.6	Problem 4: Determining Customers' Perceptions of a Chain's Quality	26
2.2.7	Problem 5: Simultaneously Benchmarking Chains and Scaling Customer Perceptions	27
2. 3	Discussion	28
2.4	Limitations and Directions for Future Research	33
2.4.1	Estimation of Variance Components	33
2.4.2	Optimization of Measurement	34
2.4.3	Substantive Areas of Interest	36
Appendix 2.1: Questionnaire		38
Appendix 2.2: Comparison of Relative Error Variance in Crossed versus Nested Designs		39
Footnotes		40
Tables		41

Chapter	Title	Page
Bibliography		48
3.	Influence of Estimation Method and Interval Estimates on the Optimality of Measurement Design	55
3.1	Methods for Estimating Variance Components	56
3.1.1	General Statistical Model	59
3.1.2	Analysis of Variance Estimation	61
3.1.3	Maximum Likelihood Estimation	62
3.1.4	Restricted Maximum Likelihood	64
3.1.5	MINQUE (Minimum Norm Quadratic Unbiased Estimation)	65
3.2	Methods to Determine Interval Estimates of Variance Components	66
3.2.1	Satterthwaite's Method	67
3.2.2	General Procedure to Construct Confidence Intervals with ANOVA Estimates	69
3.2.3	Confidence Intervals on Ratios of Variance Components	71
3.3	Empirical Illustration	74
3.3.1	Brief Description of Data Collected by Finn and Kayandé (1997)	75
3.3.2	Impact of Point Estimation Method on Optimal Designs	76
3.3.2.1	Impact on Generalizability Coefficient	76
3.3.2.2	Impact on Optimal Designs	78
3.3.3	Impact of Interval Estimation on Optimal Designs	81
3.3.3.1	Confidence Interval on the Estimate of Generalizability Coefficient	82
3.4	Discussion and Conclusions	83
Footnotes		87
Tables		88
Bibliography		94
4.	Design and Reliability Assessment of Multidimensional Measurement	97

Chapter	Title	Page
4.1	Current Methods to Assess Reliability of Multidimensional Constructs	100
4.2	Composite Scores in Marketing	101
4.3	Assessment of Generalizability of a Composite Score	103
4.3.1	Multivariate Random Effects Model	104
4.3.2	Multivariate Generalizability Coefficient	105
4.3.4	Choice of Weights	105
4.3.5	Optimization of Multi-Dimensional Measurement Designs	107
4.3.6	Estimation of Variance Covariance Components	108
4.4.	Empirical Illustration	110
4.5	Results	112
4.5.1	Univariate Results	112
4.5.2	Optimal Designs for Decision Studies	113
4.5.2.1	Problem 1: Comparison of Retail Chains	113
4.5.2.2	Problem 2: Comparison of Respondents	115
4.6	Discussion and Conclusions	119
Appendix 4.1: Questionnaire		122
Appendix 4.2: Cost Function used in the Optimization		123
Tables		124
Bibliography		127
5.	General Discussion and Conclusions	130
	Bibliography	137
Appendix I: Statistical Model Underlying Measurement in Classical Test Theory and Generalizability Theory		140
Bibliography		149
Appendix II: Questionnaire		151

List of Tables

		Page
Table 2.1:	Object of Measurement as a Function of the Purpose of Measurement	41
Table 2.2:	Analysis of Variance and Variance Component Estimates	42
Table 2.3:	Reduction of Error Variance by Sampling from Multiple Facets	43
Table 2.4:	Optimal Designs for Retail Chains as Object of Measurement, and Crossed versus Nested Designs	44
Table 2.5:	Optimal Designs for Identifying Priorities for Quality Improvement	45
Table 2.6:	Optimal Designs for Benchmarking Chains on Different Aspects of Quality	46
Table 2.7:	Optimal Designs for Simultaneously Benchmarking Quality of Chains and Determining Customers' Perceptions of Quality of any Chain	47
Table 3.1:	Comparison of Alternative Methods for Estimation of Variance Components	88
Table 3.2:	Unbalanced Data, Point Estimates, and Generalizability Coefficients for a 31 respondents, 4 aspects, 1 item Decision Study to Compare Retail Chains	89
Table 3.3:	Balanced Data, Point Estimates, and Generalizability Coefficients for a 31 respondents, 4 aspects, 1 item Decision Study to Compare Retail Chains	90
Table 3.4:	Impact of Estimation Method on Optimal Designs for Problem 1, 2, and 4 from Finn and Kayandé (1997)	91
Table 3.5:	Two-sided 90% Confidence Intervals on ANOVA Estimates of Variance Components from Balanced Data	92
Table 3.6:	Two-sided 90% Confidence Intervals on Maximum Likelihood Estimates of Variance Components from Balanced and Unbalanced Data	93
Table 4.1:	Estimated Variance-Covariance Components Matrices	124

Page

Table 4.2: Comparing 5 Retail Chains on Two Dimensions Independently, Both Dimensions Simultaneously, and a Composite of Two Dimensions 125

Table 4.3: Comparing 30 respondents on Two Dimensions Independently, Both Dimensions Simultaneously, and a Composite of Two Dimensions 126

Chapter 1

Introduction

Acquiring and interpreting information about the characteristics of objects has been a major focus of scientific inquiry in many disciplines. Such information is rarely acquired in its perfect form. Either the information acquisition process of a research study is limited to some extent, or the true state of an object's characteristic can never be known even if the information acquisition process is perfect. The information acquisition process can be limited because of several important reasons, one of which is the inability to measure the characteristic of interest under all possible conditions that affect the measurement procedure. For a simple example in the physical sciences, consider the measurement of the length of a table. A person with a ruler can take *a* measurement of the length. The measured length might depend on the person who measures the length and the particular ruler used by that person to measure the length. Thus, the measurement of the length of the table by one person with only one ruler can only be an approximation to the *true* length, if there is variation in the measure attributable to persons or rulers. On the other hand, the true state of the characteristic might be "unknowable" because we can only represent the true state isomorphically (perhaps because the characteristic is socially constructed), or because of the inherently stochastic nature of the true state. In the social sciences, we often find such a limit in the information acquisition process *and* an inability to know the true state of the object's characteristic. This combination results in a less than perfect "measurement" of the characteristic of an object.

Scientific research in social sciences has pursued the development of methods to improve the measurement of an object's characteristic, so that the true state of the characteristic is better represented, isomorphically or otherwise, by the measurement. A large number of these methods originated in education and psychology, although some were adaptations from the statistics literature. Almost all methods are statistical in nature and this thesis is limited to methods that improve the quantitative measurement and

representation of the true state of an object's characteristic, *assuming that the imperfections in measurement arise from the limits in the number of conditions under which a measurement of the characteristic can be made.* Whether a true state exists is controversial, and there are many philosophical perspectives on the issue. Although an interesting question, I abstract from it by *not assuming* the true state to be divinely determined. Instead, the true state is considered to be the "limiting value of extensive observations" (Cronbach et al. 1972) of a stable characteristic of an object, consistent with the focus of the thesis on the number of diverse conditions under which the characteristic is observed.

The academic discipline of marketing, like any other scientific discipline, has been constructed on the basis of measurement of the characteristics of objects. Firms, products, brands, consumers, markets, and distribution channels are among the objects that are of critical interest to marketing inquiry. Thus, almost every academic research study in marketing measures the characteristics of one or more of these objects, albeit in different ways. The characteristics of objects such as attitudes of consumers towards a behavior, the service quality provided by service firms, the satisfaction experienced by consumers upon use of a service or product, the image of a firm or a brand, the size of a market, market share of a brand, the demographics of a consumer population, and the choices made by consumers are commonly measured by academic researchers and practitioners in marketing. It is possible to directly observe some characteristics of some objects, e.g., choices made by consumers can be observed by scanner data. On the other hand, characteristics of many types of objects are unobservable, i.e., they cannot be observed as a fact and are therefore isomorphically represented to facilitate their measurement. For example, the satisfaction experienced by consumers or the service quality provided by firms are characteristics that are unobservable. The marketing literature has numerous scales developed ostensibly for the measurement of such characteristics, and as an indication of this activity, there are now at least two handbooks

with details of multi-item scales developed in marketing (Bruner and Hensel 1993, Bearden, Netemeyer, and Mobley 1993).

The measurement literature in marketing can be divided into the pre-1979 and post-1979 periods which differ greatly in terms of the rigor associated with the assessment of the measurement and the measure development procedure for multi-item scales. Prior to 1979, few studies provided any evidence of reliability or validity for the measures they used, with most exhibiting the folly of single indicants (Jacoby 1978). The special issue of *Journal of Marketing Research* in 1979, which contained two seminal papers by Churchill (1979) and Peter (1979), made a significant impact on the development of scales in marketing. Most marketing scales published since then have been developed on the basis of *classical test theory* methods recommended by the aforementioned seminal papers. Churchill (1979) brought together a large part of the extant literature in education and psychology to present an "updated paradigm" for measurement *scale development* in marketing. Reliability assessment was treated as an integral part of scale development. Peter (1979) focused on reliability and presented a concise summary of extant methods for measurement *scale evaluation*, either during the scale development stage or the scale application stage. However, the understanding was clear, albeit implicit, across these two papers: the recommended scale development procedures would lead to measurement scales that only need to be evaluated on their reliability when scaling respondents on some characteristic, when used for diverse managerial or academic purposes.

The origin of classical test theory methods in education and psychology has much to do with the focus of the methods on scaling a characteristic of individuals. This is so because the disciplines of education and psychology are interested primarily in individual differences. Classical test theory focuses on the ability of a measurement procedure to reliably scale a student's ability or some other characteristic relative to other students in the population. Given this focus and the seemingly parallel focus of marketing on

consumers, the adaptation of these methods in marketing was deemed to be straightforward and thought to have required no modification. Thus, measurement scales used in marketing have been evaluated for their ability to scale some characteristic of the individual respondents in a survey. However, this aspect of the evaluation has always been implicit and has almost never been explicitly discussed. The critical question that is asked in this thesis is whether scale performance when scaling a characteristic of individuals is the appropriate criterion to evaluate the information used to make various managerial decisions. Different managerial decisions require scaling of characteristics of *different* objects and therefore the appropriate criterion depends on the purpose of measurement. This thesis is an attempt not only to raise this question in a precise form but also to answer it with an alternative methodology. The process of answering this question has led us to other areas of research that are at once novel and have strong implications for managerial and academic practice of measurement.

Four specific issues given explicit attention in this thesis deserve mention at this stage. First, classical test theory methods treat error as a sample from a single undifferentiated distribution (Cronbach et al. 1972, Brennan 1983). Any study of the evolution of statistical thought will certainly accord great importance to the revolutionary thinking of Fisher (1925), who suggested that variation in a measure can arise from several controllable sources, some of which are sources of error. The impact of Fisher's work on experimental design has been considerable in marketing and there is no reason for marketing researchers to continue assuming error to be singularly attributable to a single random source in the context of evaluating the psychometric properties of a measurement procedure. In fact, the idea that variation can arise from several sources is the primary basis of the generalizability theory work presented by Cronbach and his colleagues (Cronbach et al. 1972) in educational psychology and further developed for marketing in this thesis.

Second, the thesis also suggests methods to improve the efficiency of the information acquisition process. Simply put, the data collection process for decision making can be made more efficient by taking into account the purpose of measurement. This is achieved by estimating the variation due to different sources, deciding which sources constitute noise in measurement, and so are controllable, and which source(s) constitute a signal, and then constructing a measurement instrument that leads to a better ability to measure the characteristic of interest at lowest monetary cost. These two important issues dominate the first paper of the thesis.

A multi-faceted theory of measurement requires modeling the variance in the observations as a function of the variances attributable to several sources. Such modeling is familiar in the statistics of experimental design, and results in variance components models (Searle, Casella, and McCulloch 1992). Variance component models are used throughout the thesis. There are several methods to estimate the parameters of such models and the development, description, and comparison of these methods constitute a large literature in statistics (for a selective literature review, see Khuri and Sahai, 1985 and Searle, Casella, and McCulloch 1992). The most appropriate estimation method depends on the kind of data collected by a researcher, and also has a significant impact on the implications drawn from an empirical study. In the second paper of the thesis, I have reviewed the statistical methods for estimating parameters of variance component models. I also provide an empirical comparison of the alternative methods to estimate parameters, to answer the question of whether the implications drawn from the substantive results in the first paper are dependent on the estimation method. In addition, I have presented methods, adapted from several statistical sources, to evaluate the sampling variability of the parameters estimated in a variance components model and form interval estimates of such parameters. The sampling variability can potentially have a significant impact on the implications of the empirical results, if it is such that the interval estimates have too wide a range. The empirical application of these interval

estimation methods can suggest to us whether suffficient confidence should be placed in implications drawn from the results of the first paper of the thesis.

Finally, the thesis also suggests a multivariate method that can be used to evaluate the ability of a measurement procedure to scale a *multidimensional* characteristic of an object, and also develop efficient procedures to collect information on multidimensional characteristics. Scores obtained on multiple dimensions are sometimes combined for managerial purposes to provide parsimony in information. The parsimony in information may be rewarded by a recommendation of less data collection in a decision study than would be required if one needed reliable information for each of the individual dimensions. Clearly, this reward is dependent on the covariances between the multiple dimensions. Because the parsimonious information demands less data, there could be attendant losses in reliability of information for each individual dimension. Thus, the reliability of the more diagnostic information for individual dimensions is lessened as a result. We explore the implications of such a trade-off between parsimony and diagnosticity, specifically in the context of obtaining reliable information for each objective. This multivariate extension of the concepts and methods presented in the first paper is examined in the third paper of the thesis.

Throughout the thesis, I use the example of the measurement of service quality to illustrate the conceptual issues and add a managerial context to the explanations. An additional, and perhaps more important, reason to use service quality measurement as an example is because the theory and methods are empirically illustrated in this thesis with the measurement of the service quality provided by retail chains. Thus, the common thread of service quality measurement will run through the thesis, making the context common across examples and the theory. The exception is the final paper where product quality is also measured to reflect the multidimensional nature of measuring the quality of a retail outlet. Although the focus in this thesis is on service quality for the illustration, the methods are equally applicable to many other areas of inquiry where measurement

scales are used, both in marketing and in the broader management literature. The assumption made throughout the thesis is that the characteristic(s) being measured has already been determined to be important for some managerial or academic decision-making purpose.

Bibliography

Bearden, William O., Richard G. Netemeyer, and Mary F. Mobley (1993), Handbook of Marketing Scales: Multi-Item Measures for Marketing and Consumer Behavior Research. Newbury Park, CA: Sage.

Brennan, Robert (1983), Elements of Generalizability Theory. Iowa City, Iowa: ACT Publications.

Bruner II, Gordon C. and Paul J. Hensel (1993), Marketing Scales Handbook: A Compilation of Multi-item Measures. Chicago. American Marketing Association.

Churchill, Gilbert A., Jr., (1979), "A Paradigm for Developing Better Measures of Marketing Constructs," Journal of Marketing Research, 16, (February), 64-73.

Cronbach, Lee J., Goldine C. Gleser, Harinder Nanda, and Nageswari Rajaratnam (1972) The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles. New York: John Wiley & Sons.

Fisher, R.A. (1925), Statistical Methods for Research Workers. 1st Ed., Edinburgh and London: Oliver and Boyd.

Jacoby, Jacob (1978), "Consumer Research: A State of the Art Review," Journal of Marketing, 42 (April), 87-96.

Khuri, A.I. and Hardeo Sahai (1985), "Variance Components Analysis: A Selective Literature Survey," International Statistical Review, 53, 3, 279-300.

Peter, J. Paul (1979), "Reliability: A Review of Psychometric Basics and Recent Marketing Practices," Journal of Marketing Research, 16, (February), 6-17.

Searle, Shayle R., George Casella, And Charles E. McCulloch (1992), Variance Components. New York: John Wiley and Sons.

Chapter 2

Reliability Assessment and Optimization of Marketing Measurement[1]

Criticism of the ad hoc approaches to measurement in marketing peaked 20 years ago with Jacoby's 1976 presidential address to the Association for Consumer Research. At that time few studies provided any evidence of reliability or validity for the measures they used, with most exhibiting the folly of single indicants (Jacoby 1978). The major turning point in marketing researchers' concern with measure quality was the publication of a special section on measurement in the February 1979 issue of the *Journal of Marketing Research* (Peter and Ray 1984).

The special section included two articles, Peter (1979) and Churchill (1979), that have been very influential in the development of measurement in marketing. Peter (1979) provided a comprehensive review of the traditional psychometric approach to reliability, also called classical reliability theory, and had the foresight to identify generalizability theory (Cronbach et al. 1972) as being of potential interest to marketing scholars. Churchill (1979) adopted methods already being used in the psychometric literature to present a paradigm for the development of better multi-item measures in marketing. Because of the clear prescription laid down by Churchill (1979), most scale development studies in marketing have followed the methods suggested therein, clearly resulting in a higher standard of research in marketing.

The evidence as to subsequent academic practices can be found in two projects devoted to documenting such scales. Bruner and Hensel (1993a) searched the six leading marketing journals from 1980 to 1989 for multi-item scale usage, and found 750 instances where basic psychometric information was reported. Moreover, there was a six fold increase in multi-item scale *usage* from 1980 to 1989. Their Marketing Scales

---

[1]A version of this chapter has been published in *Journal of Marketing Research*, 34 (2), 262 - 275 (with Adam Finn).

Handbook (Bruner and Hensel 1993b) describes the 588 different multi-item scales they found, classified into consumer behavior, advertising, and organizational, sales force, and miscellaneous scale sections. Bearden, Netemeyer, and Mobley (1993) report on a wider search for articles dealing with the *development* of multi-item scales in their Handbook of Marketing Scales. Consistent with the methods suggested by Churchill (1979), the structure of the description and evaluation of the scales in both Marketing Handbooks simply emulates prior handbooks of psychological scales.

Most multi-item marketing scales reported on in the Handbooks exhibit high reliability. Churchill and Peter (1984) and Bruner and Hensel (1993a) report mean alpha values of .76 and .77 respectively, whilst Peterson (1994) found the median alpha values reported in five marketing sources ranged from .76 to .81. These assessments have taken a classical reliability theory perspective, with the measures developed and then evaluated for their ability to scale respondents. In classical reliability theory, reliability is *assessed* as the degree to which a multi-item measurement instrument consistently scales a sample of individuals. Reliability can be expressed in mathematical form as,

(2.1)
$$r = \frac{\sigma^2_{\text{true score}}}{\sigma^2_{\text{observed score}}}$$

where, $\sigma^2_{\text{true score}}$ is the variation among respondents' mean scores and $\sigma^2_{\text{observed score}}$ is the sum of true score variance and error variance. Thus, a scale is seen as producing highly reliable information if it provides a consistent scaling of respondents.

Marketers have always approached scale development the same way it has been done in psychology, as if their objective is always to scale characteristics of individuals. However, many measurement applications in marketing require the scaling of objects such as firms (comparing the service quality provided by firms or the marketing orientation of firms), advertisements (comparing the effectiveness of advertisements), or brands (comparing brand image), rather than a scaling of individuals. Such scaling may require *generalization over individuals* in the population, it does not require scaling of the

individuals. Thus, a reliability assessment method that considers only the scaling of individuals is clearly limited in its scope.

Suggested refinements to scale development methods (Gerbing and Anderson 1988; Steenkamp and van Trijp 1991) have not addressed this limitation. Generalizability theory, hereafter called G-theory, explicitly addresses this limitation of classical reliability theory. G-theory, pioneered by Cronbach and his colleagues (Cronbach, Rajaratnam, and Gleser 1963; Cronbach et al. 1972), can be viewed as the most general psychometric theory, although Cooil and Rust (1994, 1995) recently presented an expected loss framework that generalizes G-theory to include non-interval scales. G-theory encompasses classical reliability theory as a special case. Moreover, the use of G-theory allows a practitioner to design efficient measurement in applied studies.

Despite frequent citation of the Peter (1979) paper, there has been little recognition of the advantages of using a G-theory approach for scale development in marketing. Rentz (1987) argued strongly for the potential value of G-theory in marketing, presenting examples based on borrowed and hypothetical data. He also illustrated the differences between classical reliability and G-theory based measures of reliability for some marketing scales using data from a small sample of students (Rentz 1988). However, there have only been three subsequent marketing citations of G-theory. Reibstein, Bateson, and Boulding (1988) recognized the multiple sources of variability when assessing the reliability of conjoint analysis. Hughes and Garrett (1988) examined the variance components for facets contributing to the observed levels of intercoder reliability in studies reported in the marketing literature. Finally, Rust and Cooil's (1994) generalization of G-theory reports tables for the levels of agreement necessary between judges to achieve levels of a proportional reduction in loss index of data quality. So while a G-theory approach to measurement has been advocated in marketing, it has not yet been adopted by marketers who are developing or using scales. Inspite of its potential, neither

academics nor practitioners have used it to ensure that money is not being wasted on unnecessary data collection during measurement for marketing decision making.

In this paper, we reiterate that the classical reliability theory perspective, which dominates scale development and academic measurement practice in marketing, is inadequate for assessing the reliability of information gathered by using a scale. Moreover, it frequently results in very inefficient solutions to managerial measurement problems. These arguments are illustrated with an empirical investigation of service quality measurement. The empirical study illustrates the importance of recognizing the purpose for which a scale is being used, and demonstrates that a generalizability approach to measurement can provide substantial savings over procedures currently advocated in the academic literature. We use the G-theory approach to design efficient measurement instruments on the basis of data collected during the development stage. The optimization takes into account both the specific purpose of proposed measurement and the monetary cost of the instrument. We next provide a discussion of the advantages of using the generalizability theory framework in marketing measurement, and conclude with limitations and directions for future research.

2.1 Using Generalizability Theory to Optimize Measurement

As opposed to the limited view of classical reliability theory, G-theory (Cronbach et al. 1972) provides a multi-faceted view of measurement, where variation in measurement arises from multiple *controllable* sources. Following the Fisherian logic of experimental design, measurement in G-theory is expressed in terms of random effects and variance components associated with each of the multiple sources of variance. Each random effect is assumed to be distributed with a mean of zero and variance equal to the variance component. The variation in the characteristic to be measured can then be expressed as the sum of the variance components associated with each source of variance.

Cronbach et al. (1972) defined a formula for the coefficient of generalizability, analogous to the reliability coefficient in classical reliability theory,

$$(2.2) \qquad E\hat{\rho}^2 = \frac{\sigma^2_{\text{universe score}}}{\sigma^2_{\text{universe score}} + \sigma^2_{\text{relative error}}}$$

where $\sigma^2_{\text{universe score}}$ is the variance component associated with any object of measurement (analogue of the true score variance in classical reliability theory) and $\sigma^2_{\text{relative error}}$ is the sum of only those variance components that affect the scaling of the levels of the object of measurement. The notation[1] is meant to show that a generalizability coefficient is "approximately equal to the expected value . . . of the squared correlation between observed scores and universe scores" (Brennan 1983, p. 17). For a study with a single facet of generalization, the coefficient of generalizability is equivalent to the reliability coefficient.

Because some of the terminology in this paper is unique to G-theory, we note some basic definitions of terms. An "object of measurement" is a factor, such as firms, advertisements, or brands, the levels of which need to be scaled by the measurement instrument. A "facet of generalization" is a factor over which the researcher requires the findings to generalize. For example, a scaling of the service quality provided by firms should generalize over respondents in the population, thus making firms the object of measurement and respondents a facet of generalization. The levels of an object or a facet are the different elements constituting the factor. Six firms would constitute 6 levels of the object of measurement 'firms' and 200 respondents would correspond to 200 levels of the facet called 'respondents'. G-theory assumes that each random effect 'i' is normally[2] distributed with mean 0 and variance σ^2_i. A generalizability study, hereafter called a G-study, is the first stage of the two-stage procedure used in G-theory in which the variance components σ^2_i associated with each effect 'i' are estimated from empirical data. The second stage, called a decision study, is an applied study, the results of which are used to make decisions. Details of G-theory are provided by Cronbach et al. (1972) and Brennan (1983). Rentz (1987) provides the most complete account in marketing; Shavelson and Webb (1991) provide a good introduction.

Scaling of the levels of an object of measurement is affected by every interaction of the object of measurement with a facet of generalization. A significant interaction between the object of measurement and a facet of generalization implies that the scaling of the levels of the object of measurement depends partially on the specific level of that particular facet of generalization. Relative error variance is the sum of the variance components associated with the interactions of the object of measurement with every facet of generalization and the random error variance. In a fully crossed study, if A is the object of measurement and B and C are facets of generalization, relative error variance can be expressed as,

$$\hat{\sigma}^2_{\text{relative error}} = \frac{\hat{\sigma}^2_{A\,X\,B}}{N_B} + \frac{\hat{\sigma}^2_{A\,X\,C}}{N_C} + \frac{\hat{\sigma}^2_{A\,X\,B\,X\,C,\ \text{random error}}}{N_B N_C} \tag{2.3}$$

where N_B and N_C are the number of levels of facets B and C respectively.

Equation 2.3 shows that the function for relative error variance is convex in the number of levels of the facets of generalization. Thus, the number of levels of the facets of generalization designed into a decision study determines the relative error variance, and therefore the expected G-coefficient.

The purpose of the measurement clearly defines the object of measurement and the associated sources of measurement error. Optimizing measurement implies identifying the most efficient allocation of resources along each of the sources that constitute error in a decision study. To structure the optimization problem, we let the decision maker choose the level of generalizability 'g' acceptable in the scaling of the object of measurement (which determines the acceptable noise in the information). The required level of generalizability will depend on the consequences that could flow from the decision study. Nunnally (1978) suggested rules of thumb such as 0.90 as the absolute minimum reliability for any applied study, with 0.95 being the desired level of reliability. The same guidelines could apply in the current context.

Given a desired generalizability, relative error variance can be reduced by sampling along the facets that contribute to error, much the same way the Spearman-Brown prophecy formula in classical reliability theory indicates that an increase in items results in higher reliability. Given 3 or 4 facets along which samples can be drawn, the question then is whether sampling should be equally distributed across facets. The answer is no, because facets contribute different amounts of error variance. Sampling should be in proportion to the size of the associated variance components. A desired level of generalizability may be achieved with several designs, each a different sampling along the facets (we call such designs iso-generalizability designs). The best choice from among these iso-generalizability designs will depend on their costs. The lowest cost design which satisfies the generalizability criterion will be the most efficient.

The optimization can be formally set up as follows. Assume the cost of measurement C is a function of the number of levels of each facet and object included in the measurement design, and the cost of each level of each facet and object. For a specific decision study, where the object of measurement is known from the managerial problem, cost C can be represented as,

(2.4) $$C = f(c_0, \tilde{C}, \tilde{N}, n_{object})$$

where c_0 is the fixed cost of the survey instrument, $\tilde{C}$ is a vector with elements representing the cost of an observation on each facet, $\tilde{N}$ is a vector with elements $n_1, n_2, \ldots, n_F$ representing the number of levels of F facets of generalization, and n_{object} is the number of levels of the object of measurement.

The cost C can be represented by different functions (such as additive or multiplicative) depending on the data collection method. The purpose here is to minimize the cost of the design subject to achieving a desired value g of the G-coefficient $E\hat{\rho}^2$. The relative error variance $\left[\hat{\sigma}^2_{\text{relative error}}(\tilde{N})\right]$ is a decreasing function in all F elements of $\tilde{N}$. Thus, the optimization problem can be formally stated as follows:

(2.5) Minimize $C = f(c_0, \tilde{C}, \tilde{N}, n_{object})$

$\tilde{N} = (n_1\ n_2\ \ldots..\ n_F)$

$\tilde{N}$ subject to,

1. $\left[E\hat{\rho}^2(\tilde{N})\right] = \dfrac{\hat{\sigma}^2_{object}}{\hat{\sigma}^2_{object} + \left[\hat{\sigma}^2_{relative\ error}(\tilde{N})\right]} \geq g$, the desired G-coefficient, and where $\hat{\sigma}^2_{object}$ and all variance components that constitute $\hat{\sigma}^2_{relative\ error}$ are known *a priori* from a G-study.
2. Each element of $\tilde{N}$ 1.
3. Each element of $\tilde{N}$ is an integer.

This integer programming problem has no analytical solution. However, it can be solved by the branch-and-bound algorithm (Salkin 1975), which is available in popular spreadsheet packages (e.g., Microsoft Excel for Windows). The branch-and-bound integer programming algorithm has been used by Sanders, Theunissen, and Baas (1989) to minimize the total number of observations (product of the number of levels of all facets of generalization) needed to achieve a desired G-coefficient. In a second paper, Sanders, Theunissen, and Baas (1991) propose a method to maximize the G-coefficient by choosing the number of observations that can be accommodated within a budget constraint. Thus, the G-coefficient is specified as a result of the optimization, not as a constraint.

Marcoulides (1995) presented an optimization method that minimizes the error variance taking into account budget constraints. Thus, his method provides for an efficient allocation of resources along each facet. He derived an analytical solution for the optimum number of levels for each facet in a multi-facet design. The number of levels of each facet in such solutions is almost always a fraction, not an integer. He then rounds these numbers to the nearest integers around the solution. Mathematically, such an approach is flawed because it does not necessarily lead to the lowest cost of all possible

integer solutions and/or the minimum error variance. Our formulation makes it clear that the chosen design will lead to at least the desired level of generalizability and will simultaneously have minimum cost of all designs that do so. Moreover, these prior optimization papers use hypothetical, not empirical variance component data.

2.2 Empirical Illustration with Service Quality Measurement

We illustrate the economic advantages of the G-theory approach with an application to the measurement of service quality. Service quality measurement was chosen for three reasons. First, service quality is an important managerial issue, so proprietary assessment studies abound (see the cases reported in Spechler 1991). Improving service quality has been identified as a key strategy for firms to profitably differentiate themselves in the marketplace (Babakus and Boller 1992; Boulding et al. 1993; Cronin and Taylor 1992; Devlin and Dong 1994; Parasuraman, Zeithaml, and Berry 1988; Rust, Zahorik, and Keiningham 1995; Zahorik and Rust 1992).

Secondly, this is an area where there is a considerable discrepancy between the common practitioner reliance on a single item scale for each aspect of service to be evaluated and the multi-item measurement scales advocated by academics. Practitioner studies commonly address managerial problems using single item scales, of unknown reliability, which seem to fit the purpose at hand (Bolton and Drew 1991a, 1991b; Devlin, Dong, and Brown 1993; Schmalensee 1994). By following the scale development paradigm put forth by Churchill (1979), Parasuraman, Zeithaml, and Berry (1985, 1988) made a significant contribution to the service quality literature in developing SERVQUAL, a multi-item measure of perceived service quality. However, SERVQUAL consists of a large number of items (at least 21), even in the shorter one-column format which directly measures perceived performance relative to expectations (Parasuraman, Zeithaml, and Berry 1994b).

Thirdly, the service quality literature is an example of an area in marketing where researchers have not recognized that the respondent is rarely the object of measurement.

From a classical reliability theory perspective, SERVQUAL provides a highly reliable scaling of respondents' perceptions of five service quality dimensions (reliability, responsiveness, assurance, empathy, and tangibles), as indicated by coefficient alphas averaging 0.88 (Parasuraman, Berry, and Zeithaml 1991). As is typical of almost all marketing measurement research, such high levels of reliability led the original developers of the scale to claim that SERVQUAL will provide reliable information for such diverse purposes as tracking the service quality provided by a firm, assessing a given firm on each of the dimensions of service quality, categorizing a firm's customers into several perceived quality segments, evaluating the level of service provided by each store in a multi-unit retail chain, and assessing a firm's service performance relative to its principal competitors (Parasuraman, Zeithaml, and Berry 1988).

However, such a generalization incorrectly assumes that the reliability of a measurement instrument[3] can be assessed independently of the purpose for which it is to be used. As evidence to the contrary, research has shown inconsistency in the SERVQUAL factor structure when it is used in different service industries (Babakus and Boller 1992; Carman 1990). Criticisms and the defense of the SERVQUAL scale development procedure (Brown, Churchill, and Peter 1993; Cronin and Taylor 1994; Parasuraman, Berry, and Zeithaml, 1993; Parasuraman, Zeithaml, and Berry 1994a; Teas 1994) have also implicitly assumed that the object of measurement in all service quality applications is respondents.

Table 2.1 lists some of the management problems that would make a retailer need to measure service quality, and shows how different problems change both the object of measurement and what constitutes the error to be generalized over.

(Insert Table 2.1 about here)

For example, the first problem is to determine what store-level factors influence the service quality provided by stores in a chain. Here the service quality provided by each store in the chain must be measured accurately, so that it can be related to the other

store-level factors. Success requires reliable measures of the service quality *for each store*. There is no need to identify which customers thought they received the best or the worst service. The measurement needs to discriminate well between stores, whilst generalizing over the perceptions of different customers. Sources of variability that affect the consistency of mean scores obtained for the stores constitute error in scaling the stores. These sources include all interactions between stores and other facets, as well as the random error component. As shown in the Table, other problems have different objects and consequent sources of error. Of course, one managerial study may address more than one of these problems.

2.2.1 Empirical Investigation

To demonstrate this approach, we conducted a generalizability study of retailer service quality, and then used the estimated variance components to identify optimal measurement designs for five different service quality measurement applications, called decision studies.

The five sources of variability, or facets, included in the G-study were, retail sectors (or type of retailer), retail chains, aspects of service quality, the items used to measure service quality, and consumers. The G-study had consumers evaluate the service quality provided by a total of nine retailers, three chosen at random from amongst the well known chains in three retail sectors. The specific chains were Eaton's, Wal-Mart and Zeller's from the department store sector, Dairy Queen, Kentucky Fried Chicken and McDonald's from the fast food sector, and Safeway, Save-on-Food, and Superstore from the grocery store sector. Each chain was evaluated on nine items, three items each for three aspects of service quality. An aspect could be any distinguishable component of the service, such as point of sale, billing, or after sales service. However, to facilitate comparisons with prior work in service quality measurement, the aspects we used were randomly chosen SERVQUAL dimensions, namely tangibles, responsiveness, and empathy. The specific items used for these aspects were randomly chosen from the

perception items used in the SERVQUAL scale (Parasuraman, Berry, and Zeithaml 1991). Because of the work undertaken to refine SERVQUAL, these items are known to be suitable measures for the respective aspects. The specific items and the response format used are shown in Appendix 2.1.

For this design, retail chains were nested within retail sectors, and items were nested within aspects. Respondents were asked to rate all nine chains on all nine items, so they were crossed with chains, sectors, items and aspects. The more fully crossed the design of a G-study, the more sources of variability of measurement can be estimated (Cronbach et al. 1972). It is possible to optimize a nested measurement design for an applied study on the basis of a fully crossed G-study, but not vice-versa.

The data were collected by mail survey in the spring of 1995. On March 20, a 12 page survey booklet was sent to a probability sample of 400 households in a Canadian city. Each household received a questionnaire with the rating questions organized into a block for each retail chain. The survey mailing and postcard follow-up produced a total of 133 responses, for an overall response rate of 35% after taking into account non-deliverable surveys. Service quality evaluations by those respondents who reported they had not dealt with a retailer over the prior 12 months were not included in the generalizability analysis. Eight respondents failed to meet this criterion for all nine retailers and so were completely eliminated from the analysis. From a traditional reliability perspective, the data provided 27 scalings of the respondents' quality perceptions - tangibles, responsiveness and empathy for the nine chains. As coefficient alpha averaged 0.88 (range 0.67 to 0.96) across the scales, the data are undoubtedly comparable with those collected in previous service quality studies.

Table 2.2 provides the traditional analysis of variance for this design, along with the associated variance components. As shown by the difference between the 10,124 potential and 6,945 actual degrees of freedom, many respondents either did not satisfy the eligibility criterion or did not provide ratings for some chains. Variance components were

estimated by the minimum variance quadratic unbiased (MIVQUE) estimation method (Hartley, Rao, and Lamotte 1978), available in SAS. Respondents, aspects, and items were assumed to be randomly chosen from their respective universes of consumers who shop at the retail chains, aspects of service quality, and items to measure aspects of service quality. We could have chosen other retail sectors, such as sporting goods or shoe stores, and other department or grocery chains for the generalizability study. Further, because the retail sectors and chains are assumed to be chosen at random, they can be replaced by others of interest in a subsequent decision application.

(Insert Table 2.2 about here)

The estimated variance components due to chains, respondents, chains by respondents, and chains by respondents by aspects, and random error were relatively high[4]. Together they accounted for about 80% of the total variance in service quality scores. The estimates of the variance components for sectors and for the interactions of sectors by aspects and of sectors by respondents by aspects were negative. As shown in Table 2.2, such estimates are typically treated as if they were zero, because all negative estimates were very close to zero (Cronbach et al. 1972). The variance component associated with a nested facet (chains or items) is confounded with the variance component due to the interaction between the nested facet and the facet in which it is nested. For example, the variance component for items is confounded with the variance component for the interaction of items and aspects, because items are nested within aspects. Such confounding is explicitly recognized in this type of analysis.

2.2.2 Reduction of Error Variance in Applied Decision Studies

The variance components from the G-study can be used to calculate the expected G-coefficients for planned decision study designs or to identify the lowest cost design for a required G-coefficient. A major advantage of G-theory is that relative error variance is partially controllable through sampling along each facet included in a decision study because it is convex in the number of levels of the facets of generalization. For example,

suppose a G-study shows the variance component associated with the interaction of retail chains with items is very small, close to zero. This means the scaling of retail chains is not dependent on items. Therefore, it would be inefficient to include more than one or two items in any future decision study in which chains are the object of measurement. However, we might expect the scaling of retail chains to vary across respondents. In such a case, increasing the number of respondents will improve the generalizability coefficient of a study in which chains are the object of measurement.

Table 2.3 quantifies the effect of increasing the number of items, aspects, and respondents on the generalizability coefficient for the service quality provided by retail chains within a retail sector. When retail chains are the object of measurement, the expected G-coefficient with just one item for one aspect and 25 respondents is about 0.75. Further, with one item per aspect, four aspects, and 35 respondents, the expected G-coefficient is about 0.90. This example illustrates how a generalizability coefficient can be forecast using the G-study data.

(Insert Table 2.3 about here)

The differential reduction in error variance achieved by sampling along multiple facets also enables an optimal design to be chosen for a decision study. In the following section, we use the general optimization procedure of Equation 2.5 to design the optimal measurement for several managerial problems.

2.2.3 Problem 1: *Benchmarking Chains Within One Retail Sector*

We first consider the benchmarking problem of comparing the service quality provided by five retail chains drawn from any one retail sector. The purpose of this optimization is to identify the lowest cost measurement design that will allow a reliable comparison of service quality provided by the retail chains. Clearly, retail chains constitute the object of measurement for the study.

Respondents are the raters of the service quality delivered by the chains, but it is expected that respondents could differ in their rank-ordering of chains. Thus, the variance

component due to the interaction between chains and respondents constitutes one source of measurement error variance. The items are all designed to measure service quality provided by each chain, but again it is expected that the rank-ordering of chains could depend on the item used to measure service quality. Thus, the variance component due to the interaction of items and chains is the second source of measurement error variance. Similarly, the variance components due to the interaction between aspects and chains, and the 3-way interaction between chains, respondents, and aspects are the other sources of error variance. Finally, random error variance, which is also confounded with several higher-order interaction variance components, constitutes another source of measurement error variance. Thus, relative error variance can be expressed mathematically as:

$$\hat{\sigma}^2_{\text{relative error}} = \frac{\hat{\sigma}^2_{c\,X\,r}}{n_r} + \frac{\hat{\sigma}^2_{c\,X\,i}}{n_i n_a} + \frac{\hat{\sigma}^2_{c\,X\,a}}{n_a} + \frac{\hat{\sigma}^2_{c\,X\,r\,X\,a}}{n_r n_a} + \frac{\hat{\sigma}^2_{\text{random error}}}{n_r n_i n_a} \tag{2.6}$$

Consistent with Equation 2.4, the cost function for the mail survey data collection method used in this study can be expressed as,

$$C = f(c_0, \tilde{C}, \tilde{N}, n_{\text{chains}}) = c_0 + c_1 N_r + c_2(n_i n_a) n_c n_s + c_3(n_i n_a n_c n_s) n_r \tag{2.7}$$

where,

1. c_0 is the fixed cost of the study.
2. c_1 is the unit cost of selecting and communicating with a respondent.
3. c_2 is the unit cost of an additional item i when designing and formatting the data collection instrument. The multiplicative term $(n_i n_a)$ is necessary because a number of items taken together constitute an aspect.
4. c_3 is the incremental cost of a lengthening of the study with an additional item on the data collection cost for each respondent.
5. c_0 = \$0, c_1 = \$5, c_2 = \$10, c_3 = \$0.20 for all problems examined in this paper.
6. n_r is the number of respondents, n_i is the number of items, and n_a is the number of aspects.
7. n_c is the number of retail chains under investigation, in this case equal to 5.

8. n_s is the number of retail sectors under investigation, in this case equal to 1.
9. N_r is the total number of respondents, equal to n_r for the crossed design and $(n_r n_c n_s)$ for the nested design.

Given this structure of the cost function, it is possible to minimize C over n_i, n_r, and n_a, subject to constraints suggested in Equation 2.5. Part I of Table 2.4 provides the optimal designs with the associated costs for a range of generalizability levels. Designs are described by the number of levels along each facet, assuming the same experimental design (crossing and nesting) as in the G-study. Note the sharp increase in costs to achieve generalizability levels of greater than 0.95. An interesting feature of these designs is that they require only one item per aspect. Technically, this result is due to the relative independence of the rating of chains and items, as reflected in the low variance component associated with interaction of chains and items (only 0.7% of total variance). Substantively, this result illustrates that the managerial preference for using a single item per aspect can in fact be optimal for certain types of problems.

Current academic research on service quality (e.g., Parasuraman, Berry, and Zeithaml 1994b) suggests benchmarking studies collect data separately from customers of each chain, nesting respondents within chains. We can calculate the variance components expected from such a nested design using the variance components from our fully crossed G-study. Thus, it is also possible to minimize C to solve for an optimal design while nesting respondents within the chains being benchmarked. Part II of Table 2.4 provides the optimal sampling and the costs associated with nested designs. Note that the number of levels required for each facet (specially, the total number of respondents N_r) are much larger for the nested designs, and so are the associated costs.

Appendix 2.2 compares mathematical expressions for relative error variance for designs when (1) respondents are crossed with chains, and (2) respondents are nested within chains. It is obvious that a nested design will always require at least as many, and in most cases, larger samples of most facets than are required for a crossed design.

Therefore, the reduction in cost with a crossed design is independent of the form of the cost function. This comparison illustrates the gain in efficiency from approaching survey measurement research from a generalizability and optimization perspective.

(Insert Table 2.4 about here)

2.2.4 Problem 2: *Identifying Priorities for Quality Improvement*

We now consider a problem that results in a different object of measurement. Suppose management is interested in determining which of five aspects of their retail chain's service quality are most in need of improvement. As this requires a reliable scaling of the five aspects of service quality, aspects become the object of measurement. The number of chains in this design is fixed at one, but we require optimal designs for comparing any five aspects for any chain within any sector. Therefore, the object of measurement becomes aspects nested within chains, which are in turn nested within sectors. The generalizability coefficient for this design can be expressed as:

$$E\hat{\rho}^2 = \frac{\hat{\sigma}^2_a + \hat{\sigma}^2_{sXa} + \hat{\sigma}^2_{cXa}}{\left(\hat{\sigma}^2_a + \hat{\sigma}^2_{sXa} + \hat{\sigma}^2_{cXa}\right) + \hat{\sigma}^2_{\text{relative error}}} \tag{2.8}$$

where,

$$\hat{\sigma}^2_{\text{relative error}} = \frac{\hat{\sigma}^2_i + \hat{\sigma}^2_{sXi} + \hat{\sigma}^2_{cXi}}{n_i} + \frac{\hat{\sigma}^2_{rXa} + \hat{\sigma}^2_{sXrXa} + \hat{\sigma}^2_{cXrXa}}{n_r} +$$

$$+ \frac{\hat{\sigma}^2_{rXi} + \hat{\sigma}^2_{sXrXi} + \hat{\sigma}^2_e}{n_r n_i}$$

As shown in Equation 2.8, relative error variance can be reduced by sampling more respondents and/or increasing the number of items. Table 2.5 provides optimal designs and the associated costs for generalizability levels ranging from 0.80 to 0.95. The cost function is identical to that in Equation 2.7, with number of aspects fixed at 5 and number of chains and sectors fixed at 1. Note that the number of items required for this problem is much higher than for Problem 1. This clearly illustrates the influence of the purpose of measurement on the optimality of a measurement design.

The stringent requirements for number of respondents (64) and number of items (8) per aspect to achieve generalizability of 0.90 probably results from the fact that SERVQUAL was not developed to distinguish between aspects, rather it was developed to distinguish between perceptions of respondents. Although Parasuraman, Zeithaml, and Berry (1988) specifically claim SERVQUAL can provide reliable information for this issue, SERVQUAL items were originally selected and have since been refined for their ability to scale respondents. Thus, this problem illustrates an inadequacy of classical reliability theory methods.

(Insert Table 2.5 about here)

2.2.5 Problem 3: *Benchmarking Chains on Different Aspects of Quality*

In this third problem, the retail chain's management is interested in the more detailed problem of benchmarking how competing chains are doing on different aspects of service quality. For this problem, the measurement procedure should be able to reliably scale chains on the different aspects. Thus, the object of measurement becomes the interaction between chains and aspects. All interactions with three or more facets that also include chains and aspects will contribute, along with random error, to the relative error for this problem. Table 2.6 provides optimal designs for this problem, which also uses the same cost function as in Equation 2.7. Note the increase in the number of observations required relative to Problem 2 for each level of the G-coefficient, because of the more detailed nature of this problem.

(Insert Table 2.6 about here)

2.2.6 Problem 4: *Determining Customers' Perceptions of a Chain's Quality*

The academic literature on service quality measurement has focused on scaling customers' perceptions of the quality of a specific service firm, and has reported high reliability for this scaling of respondents nested within firm as the object of measurement. A reliable scaling of respondents could be used to "categorize a firm's customers into several perceived-quality segments (Parasuraman, Zeithaml, and Berry 1988, p. 35).

Although this is a less common managerial problem, the method presented herein should result in a high G-coefficient for respondents as object of measurement, with the numbers of items and aspects commonly used in the literature and a design with respondents nested within chain. The G-coefficient for this design can be mathematically expressed as follows:

$$E\hat{\rho}^2 = \frac{\hat{\sigma}^2_r + \hat{\sigma}^2_{sXr} + \hat{\sigma}^2_{cXr}}{\left(\hat{\sigma}^2_r + \hat{\sigma}^2_{sXr} + \hat{\sigma}^2_{cXr}\right) + \hat{\sigma}^2_{\text{relative error}}} \tag{2.9}$$

where,

$$\hat{\sigma}^2_{\text{relative error}} = \frac{\hat{\sigma}^2_{rXa} + \hat{\sigma}^2_{sXrXa} + \hat{\sigma}^2_{cXrXa}}{n_a} + \frac{\hat{\sigma}^2_{rXi} + \hat{\sigma}^2_{sXrXi} + \hat{\sigma}^2_e}{n_a n_i}$$

The predicted G-coefficient for 4 items within each of 5 aspects (making a 20 item scale) is 0.89, which is close to the average of Cronbach's alpha 0.88 obtained for the SERVQUAL scale. Although interesting, this result is not surprising because of the nature of the generalizability coefficient, which can be viewed in this empirical context as the expected value of the usual reliability coefficient across retail sectors, retail chains, aspects and items.

Using the optimization framework, the optimal design to achieve a G-coefficient of 0.95 is 1 item each for 19 aspects. If we accept the current service quality measurement suggestion of 5 aspects, we can impose an additional constraint on the number of aspects to be less than or equal to 5. To achieve a G-coefficient of 0.90, the measurement study would require 6 items for each of 5 aspects. However, it becomes impossible to achieve a G-coefficient of 0.95 because of the constraint on the number of aspects (see Equation 2.9). This result is independent of the cost function used in the optimization.

2.2.7 Problem 5: *Simultaneously Benchmarking Chains and Scaling Customer Perceptions*

If the purpose of a managerial study is to simultaneously provide a solution to two or more different managerial problems, then the optimization can be easily structured to

account for such a situation. Further, there might be significant cost savings in the process. For example, consider the multiple problems of benchmarking the service quality of retail chains within a retail sector (Problem 1, Part II) and determining consumer perceptions of a chain's quality (Problem 4). Assuming a design with respondents nested within retail chain, we can formalize a single problem, the optimization of which will provide a simultaneous optimal solution to each. Such a solution can be, but is not necessarily, an optimal solution to each individual problem.

This example is formalized to derive solutions for both problems. The expressions for G-coefficients for both problems enter the constraint set, along with two other constraints imposed for this new problem. First, because respondents are being compared, it is necessary to set a minimum number of respondents to be compared. Second, because chains are being compared, it is also necessary to set the minimum number of chains to be compared. The minimum number of respondents was set at 30, while the number of chains was set at 5. Table 2.7 shows the optimal designs for achieving a reliability of 0.90 for the solution of both problems. Row 1 of the table shows the design for comparing chains only. The number of respondents required is 40 whilst the number of aspects is 6. Now examine row 2, which provides the optimal design for comparing respondents only. The number of respondents required remains at 30, whilst the number of aspects required is now 9. Thus, one solution (40 respondents, 1 item, 9 aspects, and 5 chains) that will satisfy all constraints is shown in row 3. However, this solution does not minimize the cost at the same time. Row 4 show the optimal solution for the problem. Note that the individual solution for each problem does not satisfy the constraint for the other problem.

(Insert Table 2.7 about here)

2.3 Discussion

The application of G-theory to the measurement of service quality in the current study illustrates a general method for designing measurement instruments so that subsequent studies can be optimal in terms of sampling requirements, psychometric

standards, and the cost of measurement. The optimization is primarily based on the identification of the object of measurement, which is dependent on the problem being examined. If the variance components attributable to several sources of variance are known from a G-study and the object of measurement is defined, it is possible to optimize the design in terms of the number of levels of each facet. We demonstrated an integer programming approach where the cost of measurement was minimized subject to a constraint of a pre-specified G-coefficient. Alternatively, we can determine the G-coefficient resulting from different sampling designs and choose a specific design with advance knowledge of the expected generalizability and associated cost of the measurement. If managers face a budget constraint for a decision study, then the appropriate optimization formulation would be to maximize the G-coefficient subject to the budget constraint.

G-theory has great potential for optimizing measurement, but has not found acceptance in marketing. One concern may have been the cost of carrying out a two-stage study. However, the cost of a G-study can be more than offset by the savings from a single decision study. Using the cost function in Equation 2.7, the cost of our G-study was $4972. The cost of achieving a G-coefficient of .95 in a decision study that optimally compares the service quality of 5 chains within a retail sector (Problem 1 above) was estimated to be $1166 (see Table 2.4). Thus, the total cost of the two-stage study is only $6138. There are no specific recommendations within current service quality measurement procedures about the number of respondents to include in such a decision study. However, previous researchers have used approximately 4 items for each of 5 aspects and about 200 respondents per chain. This results in a G-coefficient of 0.95, but at a cost of $10000. Thus, the two-stage procedure suggested in this paper is significantly less costly than the procedure implicitly being advocated in the current literature. This illustrates the economic benefits of using the G-theory framework to optimize the design of measurement.

The measurement literature in marketing has followed Churchill's (1979) scale development paradigm and devoted a considerable amount of research to assessing and trying to improve the reliability and validity of multi-item scales. However, the literature has largely ignored the fact that there are many different underlying reasons for conducting a measurement study. If these differences had no bearing on the measurement procedure, ignoring them would be acceptable. However, as demonstrated in the empirical study of service quality, the underlying purpose of measurement determines the object(s) of measurement and therefore cannot be ignored in the assessment and design of the measurement procedure.

Current measurement research in marketing has consistently, if only implicitly, treated customers as the object of measurement. It has ignored the possibility that other facets, such as outlets, firms, occasions, or employees, not customers, may be the object of measurement. Most managerial purposes require generalization over individuals, whereas most measurement research in marketing fails to do so and ignores other facets. For example, Brown and Swartz (1989) collected data from customers of 13 physicians to examine the applicability of a SERVQUAL type measurement procedure to professional services. However, when investigating the reliability of the measurement instrument in scaling respondents, they ignored the effect of potential differences between physicians. Our paper makes an advance over such measurement procedures by explicitly recognizing, in this example, the existence of interactions between physicians and scale items, which can potentially result in differences in scale factor structure across physicians.

Marketing academics have not always recognized the theoretical importance of these differences in objectives. In the service quality literature, one consequence has been confusion about its relationship with customer satisfaction (see Iacobucci, Grayson, and Ostrom 1994 for a useful review), because they become virtually indistinguishable if the respondent is the object of measurement (Bitner and Hubbert 1994). Surprisingly, there

has been greater recognition of this in customer satisfaction literature, where Hauser, Simester, and Wernerfelt (1994) advocated managerial use of evaluation on several facets, and Fornell (1992) and his associates (Anderson, Fornell, and Lehmann 1994; Anderson 1994) have developed and used aggregate measures of satisfaction.

As shown in our examples, the generalizability coefficient for a specific design varies with the source of variance that becomes the object of measurement. In our case, more than acceptable levels of generalizability (.95) could be obtained with a single item per aspect, 7 aspects and as few as 68 respondents when benchmarking the service quality of retail chains within the same retail sector. This efficiency was achieved by recognizing the advantage of using a design crossing retail chains and respondents. This application clearly demonstrates there are instances where the apparent practitioner preference for measuring quality with a single item per aspect is well justified. Moreover, the results show that if respondents have sufficient experience with the alternatives to be able to evaluate more than one, a crossed design substantially reduces the cost of a benchmarking study. It is to be emphasized, though, that there are situations where crossed designs may not be feasible and therefore the researcher may be constrained to use a nested design. For example, evaluating multiple outlets of a fast food chain across different geographic regions might be best achieved with a nested design because of the high probability that consumers will not have had any experience of outlets in the foreign region.

When using the G-theory approach, the researcher must clearly specify the domain from which to sample, as this is the domain over which the information from a measurement procedure is expected to generalize. Variance components and G-coefficients may not generalize beyond the specified domain. The substantive results reported in this paper only apply to the domain investigated in the G-study. The optimization results cannot just be assumed to extend to any service, in any country, when measured with any items. They apply to the universe of retail sectors, because we selected from amongst retail sectors when conducting the G-study. We did not explore

the possibility of a systematic influence of country on service quality scores. Had country been another facet in the G-study (obviously necessitating a more nested design), we could draw conclusions about optimal measurement designs for retail firms in other countries. The results apply to the universe of SERVQUAL perception items, because we selected from amongst them when conducting the G-study.

In predictions of the generalizability coefficient which will be obtained in a decision study, G-theory assumes that the levels are chosen at random for each of the facets over which generalization is to be made. G-theory does not identify some levels of a facet of generalization as better than others. Thus, it does not identify which of the three or four items (aspects) used in a G-study are the best one or two items (aspects) to select for use in the decision study.

In addition, G-theory assumes that variance components remain stable from generalizability to decision studies. The assumption of stability is common to measurement research based on classical reliability theory. This assumption is critical for the optimization of measurement to remain valid. We speculate that only major changes in a market will cause instability in variance components. But it is also possible to verify the assumptions made in the optimization exercise by reexamining the variance components and the G-coefficient in the decision study.

To summarize, current reliability assessment *methods* lack validity when a measurement instrument is used for different purposes. Neither are the methods efficient, because the cost of measurement can become unnecessarily high. In addition, the expected loss framework suggested by Rust and Cooil (1994) shows that the reliability in measurement is inversely related to the loss expected from using the information for decision making. A measurement instrument that has not been evaluated for a specific purpose might be used for that purpose, thus resulting in a higher loss than expected. Rust and Cooil's framework assesses the loss expected from using the information in the form

of a 0 to 1 index. We extend their framework to suggest that the monetary cost of measurement can be minimized simultaneously with the expected loss.

Thus, the major contribution of this paper is to demonstrate the value of a method that can be used to control the quality of information produced in measurement applications. And, most importantly, this *method* is generalizable across concepts and measurement application contexts.

2.4 Limitations and Directions for Future Research

2.4.1 Estimation of Variance Components

Cronbach et al. (1972) expressed apprehension that the G-theory approach might not find acceptance in the research community because of its potential complexity. Most published studies that have used generalizability theory have been in the area of educational testing, have 2-3 factors with small number of levels per factor, and generally obtain balanced data. However, marketing applications would typically need to handle larger number of factors with larger numbers of levels, and unbalanced data, which are to be expected because of missing observations in survey research. This added complexity places greater demands on methods of estimation and statistical inference. Balanced data simplify the estimation of variance components and their confidence intervals (for an introduction to the latter, see Burdick and Graybill 1992). Unbalanced data lead to several estimation problems, such as the confounding of estimates of variance components, biased estimates, greater likelihood of negative estimates, and excessive computational requirements. In addition, it is more difficult to derive expressions for confidence intervals around variance components estimated from unbalanced data. One clear recommendation for unbalanced data is to avoid the estimation of variance components by the method of analysis of variance. Fortunately, most of these estimation problems can be overcome by criterion-based estimation methods such as MINQUE (Rao 1971), of which MIVQUE is a special case, and maximum likelihood methods (Hartley and Rao 1967), so they should not necessarily limit the use of G-theory for marketing research.

A negative estimate of a variance component is a particularly troubling problem in the application of G-theory, because it can occur even with balanced data. In such cases, Cronbach et al. (1972) and Brennan (1983) recommend setting the negative estimate to zero. They differ on the procedures to adjust the other variance components after setting a negative variance estimate to zero. Cronbach et al. (1972) suggest that the other variance components should not be adjusted in order to maintain unbiasedness of estimates. Brennan (1983) suggests recalculating the variance estimates adjusting for the negative variance estimates set to zero. Alternatively, if some distributional assumptions are made, Bayesian procedures prevent the problem of negative variance estimates (Brennan 1983). In addition, maximum likelihood methods constrain the estimates to be non-negative. Constraining an estimate to be non-negative implies that other estimates in the model could be biased. Further, these alternative methods are computationally burdensome for models that include a large number of levels for a factor such as respondents, common in marketing research. Applications of G-theory to marketing measurement would need advances in estimation and statistical inference methods that can overcome the problems that arise with unbalanced data and large number of levels of factors. Because a detailed discussion of the estimation and issues in variance components models is beyond the scope of this paper, we refer the interested reader to comprehensive reviews of statistical issues in variance component models by Khuri and Sahai (1985) and Searle, Casella, and McCulloch (1992).

2.4.2 Optimization of Measurement

Two important assumptions are made in this paper with respect to the optimization. First, the optimizations assume there will be no missing data in the decision studies. This strong assumption is likely to be violated often in survey research. One way to relax this assumption when generating optimal designs may be to assume the distribution of missing cells across different facets in the generalizability study will hold for the decision study, and then generate optimal designs taking it into account. Thus, the

number of levels of an individual facet will increase in proportion to the expectation of missing data on that facet. In addition, it would be of interest to determine whether there is a pattern of differences in the distribution of missing data between nested and crossed designs. Such differences could be taken into account in the optimization of measurement designs.

Secondly, the optimization simply identifies the number of respondents required, ignoring the question of response rate. Our numerical results assume the response rate remains constant with an increase in the number of levels of any facet, including the object of measurement. It may be possible to represent the survey response rate as a nonlinear function of the number of items. Essentially, this is an additional penalty for increasing the length of the questionnaire and might have a significant impact on the optimal designs.

Statistical sampling is one way to generate optimal designs. Statistical design issues, such as whether a facet(s) is crossed or nested, and fixed or random, also determine the optimal design. In addition, the efficiency of a measurement design will depend on the form of data collection and the format of the stimulus material used in the study if the layout of a questionnaire affects the variance components in a systematic manner. For example, instead of putting all items for a chain in one block, we could have created a questionnaire with a block of chains for each item. If respondents respond to the stimulus by first setting a mean for a block and then distributing ratings around the mean, there could be predictable changes in the variance components due to items and chains. If the purpose of measurement is to differentiate between chains, the optimal format may not be the same as if the purpose is to differentiate between items. Thus, the layout of the questionnaire could be used to make a scale more generalizable. Similarly, alternative data collection methods (e.g., questionnaires, telephone surveys, computer interactive surveys, etc.) will not only have different functional forms for their cost, they might be expected to generate different variance components, resulting in different optimal

measurement designs. The interesting research question is the extent to which these methods systematically influence the generalizability of a scale.

2.4.3 Substantive Areas of Interest

We do not consider service quality measurement to be unique in its shortcomings. We suspect similar results and measurement efficiency gains may be achievable in other marketing areas, ranging from the obvious, such as customer satisfaction, to market orientation (Narver and Slater 1990), advertising effectiveness (Lastovicka 1983) and brand equity (Aaker 1996, Ch. 10). For example, developing a new scale for comparing the effectiveness of advertisements requires an analysis that treats advertisements as the object of measurement and respondents as a facet of generalization. Thus, the G-theory framework has the potential to impact measurement practices in several substantive areas in marketing.

The central theme of the current paper is that measurement is influenced by the conditions under which it is conducted. The generalizability of the information collected by a measurement procedure is dependent on the number and kind of conditions under which the characteristic is measured. The G-theory method is similar to meta-analysis in the sense that both explore such dependencies in information. A meta-analysis most commonly investigates the generalizability of the relationship between two variables, by estimating the influence of inter-study differences on the relationship in an analysis of variance framework (Farley and Lehmann 1986). The G-theory method investigates the generalizability of information about the characteristic of an object, by estimating the influence of the conditions under which the measurement is conducted.

The similarity between meta-analysis and G-theory leads to the recognition of one immediate avenue of future substantive work. Some service quality researchers have conducted relatively programmatic research on service quality. Data collected within some or even across several of their studies could be re-analyzed in a G-study estimating the variance components associated with different sources of variance. Then, in the

absence of context specific data, these variance component estimates could be used to approximate optimal measurement for a proposed decision study. Thus, the volumes of data collected by the service quality research community could be used for gaining further insights in this important area of research. Such a cross-study generalizability analysis would be quite similar to a meta-analysis. This would also provide an additional type of empirical generalization in marketing (see the 1995 special issue of *Marketing Science* on Empirical Generalizations). Similar analyses could be conducted across other substantive areas in marketing.

Appendix 2.1

Directions: The following statements ask how you feel about the service provided by some XYZ area department store chains, grocery store chains, and fast-food chains. Please indicate the extent of your agreement with each statement about each chain. Circle a '10' if you very strongly agree, and circle a '0' if you very strongly disagree. If your feelings lie between these two extremes, circle a number in between '10' and '0' that best shows your level of agreement. There are no right or wrong answers- we are interested in your views of the service provided by the chains.

The following statements are about Eaton's department store chain.

1. Eaton's stores are visually attractive.
2. Eaton's employees appear neat and tidy.
3. Eaton's promotional materials are visually appealing.
4. Eaton's employees give you prompt service.
5. Eaton's employees are always willing to help you.
6. Eaton's employees are never too busy to respond to your requests.
7. Eaton's employees give you personal attention.
8. Eaton's employees have your best interests at heart.
9. Eaton's employees understand your specific needs.

Note: Each statement was accompanied by an 11-point scale anchored at the end-points by the labels "Very Strongly Disagree" (= 0) and "Very Strongly Agree (= 10). The intermediate scale points were not labeled. Also, the statements were not numbered.

Appendix 2.2

Comparison of Relative Error Variance in Crossed versus Nested Designs

For a design with respondents crossed with chains,

$$(A2.1) \qquad \hat{\sigma}^2_{\text{relative error}} = \frac{\hat{\sigma}^2_{cXr}}{n_r} + \frac{\hat{\sigma}^2_{cXi}}{n_i n_a} + \frac{\hat{\sigma}^2_{cXa}}{n_a} + \frac{\hat{\sigma}^2_{cXrXa}}{n_r n_a} + \frac{\hat{\sigma}^2_{\text{random error}}}{n_r n_i n_a}$$

For a design with respondents nested within chains,

$$(A2.2) \qquad \hat{\sigma}^2_{\text{relative error}} = \frac{\hat{\sigma}^2_r + \hat{\sigma}^2_{cXr} + \hat{\sigma}^2_{sXr}}{n_r} + \frac{\hat{\sigma}^2_{cXi}}{n_i n_a} + \frac{\hat{\sigma}^2_{cXa}}{n_a} +$$

$$+ \frac{\hat{\sigma}^2_{rXa} + \hat{\sigma}^2_{rXaXs} + \hat{\sigma}^2_{cXrXa}}{n_r n_a} + \frac{\hat{\sigma}^2_{rXi} + \hat{\sigma}^2_{rXiXs}}{n_r n_i n_a} + \frac{\hat{\sigma}^2_{\text{random error}}}{n_r n_i n_a}$$

where, r stands for respondents, c for chains, s for retail sectors, i for items, and a for aspects.

Note that the expression for the crossed design (A1) is a restricted version of the expression for the nested design (A2). Therefore, a nested design will always have at least as much relative error variance as a crossed design. Moreover, terms with variance due to respondents enter the nested design expression for relative error variance, and such terms will almost always be positive and among the highest in a service quality study. In such cases, relative error variance in a nested design will be much higher than that in a crossed design, implying therefore that greater sampling is required to attain the same level of generalizability.

Footnotes

1. The generalizability coefficient is denoted by $E\hat{\rho}^2$. This notation is simple, yet imprecise because the generalizability coefficient is an estimate of the expected value of ρ^2. Thus, $E\hat{\rho}^2$ is not the expectation of the estimate of ρ^2, but the estimate of the expected value of ρ^2. It is used for simplicity, and should not be misunderstood to imply the expectation of an estimate.

2. The assumption of normality is not a strict requirement of G-theory. However, estimation and statistical inferences are simplified with an assumption of normality.

3. In the strictest sense, reliability is not a characteristic of a scale or instrument, it is a characteristic of the data or information gathered by using the scale. The same scale can produce data that are reliable and other data that are unreliable. However, because the term "reliability of a scale or measurement instrument" is commonly used in marketing research, we use it loosely in this paper to make arguments. When we use the term reliability of a scale, it should be interpreted as the reliability of information gathered by using the scale.

4. The unbalanced nature of the data in our study does not allow for straightforward statistical tests on the variance components. Therefore, for the purpose of statistical inference only, we used a subset of 65 respondents who provided responses to every item in the questionnaire, to derive the variance components and the associated F-statistics. This subset, consisting of balanced data, accounted for 76% of the full unbalanced data; estimates of all variance components were similar in relative size to the estimates from the full unbalanced data. All estimates greater than zero are significantly different from zero (ProbF < 0.05), except the estimate of the variance component associated with the interaction of sectors and respondents. All negative estimates are not significantly different from zero. Some F-tests are approximate and derived using Satterthwaite's (1946) method.

Table 2.1
Object of Measurement as a Function of the Purpose of Measurement

Purpose of Measurement/ Underlying Management Problem	Measurement Need	(Object of Measurement) Scaling Of	(Facets of Generalization) Generalize Over
1. Determine the relationship between store characteristics and service quality of stores	The service quality being provided by each of the stores in the chain	Stores in the chain	Customers, aspects, items
2. Are customers being lost because our competitors are providing better service quality?	Identify how our chain's service quality compares with the service quality of our competitors	Competitors within an industry	Stores, customers, aspects, items
3. What areas of our business activity are most in need of management attention?	Identify what aspects of our service are most in need of improvement	Aspects of business activity	Stores, customers, items
4. How well are we doing on different aspects of our business activity relative to our competitors?	Identify the performance of our chain on different aspects relative to the performance of competitors on the same aspects	Aspects of the business activity of competitors within an industry	Customers, items
5. Determine whether those customers who pay full price are getting better service	Determine the service quality perceptions of different categories of customers	Customers	Stores, aspects, items

Table 2.2
Analysis of Variance and Variance Component Estimates

Source of Variation	Potential df	Actual df	Sums of Squares	Mean Square	Variance Component	Percent of Total Variance
Chains	6	6	3437.4	572.89	0.694	14.17
Retail Sectors	2	2	558.5	279.27	0.000	0.00
Respondents	124	124	8091.4	65.25	0.850	17.23
Aspects	2	2	1056.0	528.02	0.193	3.91
Scale items	6	6	400.0	66.67	0.074	1.45
Retail Sectors by Respondents	248	229	2832.4	12.37	0.016	0.32
Retail Sectors by Aspects	4	4	72.6	18.14	0.000	0.00
Retail Sectors by Items	12	12	101.9	8.49	0.018	0.36
Chains by Respondents	744	416	5007.6	12.04	1.053	21.34
Chains by Aspects	12	12	354.1	29.51	0.093	1.89
Chains by Items	36	36	132.2	3.67	0.033	0.68
Respondents by Aspects	248	248	2088.7	8.42	0.260	5.27
Respondents by Items	744	742	1465.2	1.97	0.158	3.20
Retail Sectors by Respondents by Aspects	496	457	1211.7	2.65	0.000	0.00
Retail Sectors by Respondents by Items	1488	1363	1330.6	0.98	0.107	2.16
Chains by Respondents by Aspects	1488	827	2178.6	2.63	0.633	12.82
Error	4464	2459	1848.3	0.75	0.751	15.23
Total	10,124	6,945	32167.3		4.934	100.00

Table 2.3
Reduction of Error Variance by Sampling from Multiple Facets[a]

	G-Study	Alternative Decision Studies			
$n_{sectors}$ =	1	1	1	1	1
$n_{respondents}$ =	1	25	35	50	100
$n_{aspects}$ =	1	1	4	4	4
n_{items} =	1	1	1	1	1
Source of Variation	Estimate	Expected Variance Component[b]			
retail chain	0.69	0.69	0.69	0.69	0.69
retail sector	0.00	0.00	0.00	0.00	0.00
respondents	0.85	0.03	0.02	0.02	0.01
aspects	0.19	0.19	0.05	0.05	0.05
items[c]	0.07	0.07	0.02	0.02	0.02
sector by respondents	0.02	0.00	0.00	0.00	0.00
sector by aspects	0.00	0.00	0.00	0.00	0.00
sector by items	0.02	0.02	0.00	0.00	0.00
chain by respondents[d]	*1.05*	*0.04*	*0.03*	*0.02*	*0.01*
chain by aspects	*0.09*	*0.09*	*0.02*	*0.02*	*0.02*
chain by items	*0.03*	*0.03*	*0.01*	*0.01*	*0.01*
respondents by aspects	0.26	0.01	0.00	0.00	0.00
respondents by item	0.16	0.01	0.00	0.00	0.00
sector by respondents by aspects	0.00	0.00	0.00	0.00	0.00
sector by respondents by items	0.11	0.00	0.00	0.00	0.00
chain by respondents by aspects	*0.63*	*0.03*	*0.00*	*0.00*	*0.00*
error	*0.75*	*0.03*	*0.01*	*0.00*	*0.00*
Relative error variance	2.56	0.22	0.07	0.06	0.05
G-coefficient for chains	0.21	0.76	0.91	0.92	0.94

[a] All numbers truncated to 2 decimal places. Thus, a variance component of 0.00 does not necessarily imply that it is zero.

[b] Represents the estimated variance component for the random effect divided by the number of levels of the factors (other than the object of measurement) in the random effect.

[c] Because items are nested within aspects, all variance components that include items are divided by the number of items *and* the number of aspects.

[d] The variance components for the interactions of the object of measurement (chains) with the *facets of generalization* are in italics.

Table 2.4
Optimal Designs for Retail Chains as Object of Measurement, and Crossed versus Nested designs

I. Customers Crossed With Chains								II. Customers Nested Within Chains					
		Facet Sampling						Facet Sampling					
$E\rho^2$	σ^2_{rel}	n_i	n_a	n_r	N_r[a]	N_t[b]	Cost ($)	n_i	n_a	n_r	N_r	N_t	Cost ($)
0.80	0.174	1	2	16	16	160	212	1	4	17	85	340	693
0.81	0.163	1	2	18	18	180	226	1	4	19	95	380	751
0.82	0.152	1	2	20	20	200	240	1	4	20	100	400	780
0.83	0.142	1	2	23	23	230	261	1	4	22	110	440	838
0.84	0.132	1	2	26	26	260	282	1	4	24	120	480	896
0.85	0.123	1	3	19	19	285	302	1	5	24	120	600	970
0.86	0.113	1	3	22	22	330	326	1	5	27	135	675	1060
0.87	0.104	1	3	25	25	375	350	1	5	30	150	750	1150
0.88	0.095	1	3	29	29	435	382	1	6	31	155	930	1261
0.89	0.086	1	3	35	35	525	430	1	6	35	175	1050	1385
0.90	0.077	1	4	31	31	620	479	1	6	40	200	1200	1540
0.91	0.069	1	4	38	38	760	542	1	8	41	205	1640	1753
0.92	0.060	1	5	38	38	950	630	1	7	52	260	1820	2014
0.93	0.052	1	5	50	50	1250	750	1	9	56	280	2520	2354
0.94	0.044	1	6	56	56	1680	916	1	10	67	335	3350	2845
0.95	0.037	1	7	68	68	2380	1166	1	12	80	400	4800	3560
0.96	0.029	1	9	82	82	3690	1598	1	15	100	500	7500	4750
0.97	0.021	1	13	99	99	6435	2432	1	17	145	725	12325	6940
0.98	0.014	1	18	159	159	14310	4557	1	24	225	1125	27000	12225
0.99	0.007	1	37	304	304	56240	14618	1	44	475	2375	104500	34975

[a] Total Number of Respondents.
[b] Total Number of Observations = n_{items} X $n_{aspects}$ X $n_{respondents}$ X n_{chains}, where n_{chains} for this problem is 5.

Table 2.5
Optimal Designs for Identifying Priorities for Quality Improvement

$E\rho^2$	σ^2_{rel}	n_i per n_s	n_r	Cost ($)
0.80	0.072	4	29	461
0.85	0.051	5	43	680
0.90	0.032	8	64	1232
0.95	0.015	16	132	3572

Table 2.6
Optimal Designs for Benchmarking Chains on Different Aspects of Quality

$E\rho^2$	σ^2_{rel}	n_i per n_a	n_r	Cost ($)
0.80	0.0233	3	73	2210
0.85	0.0165	4	102	3550
0.90	0.0104	6	159	7065
0.95	0.0049	13	296	23970

Table 2.7
Optimal Designs for Simultaneously Benchmarking Quality of Chains and Determining Customers' Perceptions of the Quality of any Chain

Problem No.	Object of Measurement	n_c	n_r	n_a	n_i	$N_t = n_c n_r n_a n_i$	Cost ($)	Comments
(1)	Chains only	5	40	6	1	1200	1540	Does not satisfy (4)
(4)	Respondents only	5	30	9	1	1350	1470	Does not satisfy (1)
(1 & 4)	Chains and respondents	5	40	9	1	1800	1810	Satisfies both, but not optimal
(1 & 4)	Chains and respondents	5	34	9	1	1530	1606	Satisfies both and is optimal

Bibliography

Aaker, David A. (1996), Building Strong Brands. New York: Free Press

Anderson, Eugene W. (1994), "Cross-Category Variation in Customer Satisfaction and Retention," Marketing Letters, 5 (January), 19-30.

_____, Claes Fornell and Donald R. Lehmann (1994), "Customer Satisfaction, Market Share and Profitability: Findings from Sweden," Journal of Marketing, 58 (July), 53-66.

Babakus, Emin and Gregory W. Boller (1992), "An Empirical Assessment of the SERVQUAL Scale," Journal of Business Research, 24, 253-268.

Bearden, William O., Richard G. Netemeyer, and Mary F. Mobley (1993), Handbook of Marketing Scales: Multi-Item Measures for Marketing and Consumer Behavior Research. Newbury Park, CA: Sage.

Bitner, Mary Jo and Amy R. Hubbert (1994), "Encounter Satisfaction Versus Overall Satisfaction Versus Quality: The Customers Voice," in Service Quality: New Directions in Theory and Practice, Roland T. Rust and Richard L. Oliver, eds. Thousand Oaks, Ca.: Sage Publications.

Bolton, Ruth N. and James H. Drew (1991a), "A Longitudinal Analysis of the Impact of Service Changes on Customer Attitudes," Journal of Marketing, 55 (January), 1-9.

_____ and James H. Drew (1991b), "A Multistage Model of Customers Assessments of Service Quality and Value," Journal of Consumer Research, 17 (March), 375-384.

Boulding, William, Ajay Kalra, Richard Staelin and Valarie A. Zeithaml (1993), "A Dynamic Process Model of Service Quality: From Expectations to Behavioral Intentions," Journal of Marketing Research, 30 (February), 7-27.

Brennan, Robert (1983), Elements of Generalizability Theory. Iowa City, Iowa: ACT Publications.

Brown, Stephen W. and Teresa A. Swartz (1989), "A Gap Analysis of Professional Service Quality," Journal of Marketing, 53 (April), 92-98.

Brown, Tom J., Gilbert A. Churchill, Jr., and J. Paul Peter (1993), "Research Note: Improving the Measurement of Service Quality," Journal of Retailing, 69 (Spring), 127-139.

Bruner II, Gordon C. and Paul J. Hensel (1993a), "Multi-item Scale Usage in Marketing Journals: 1980 to 1989," Journal of the Academy of Marketing Science, 21 (Fall), 339-344.

_____ and _____ (1993b), Marketing Scales Handbook: A Compilation of Multi-item Measures. Chicago. American Marketing Association.

Burdick, Richard K and Franklin A. Graybill (1992), Confidence Intervals on Variance Components. New York: Marcel Dekker.

Carman, James M. (1990), "Consumer Perceptions of Service Quality: An Assessment of the SERVQUAL Dimensions," Journal of Retailing, 66 (Spring), 33-55.

Churchill, Gilbert A., Jr., (1979), "A Paradigm for Developing Better Measures of Marketing Constructs," Journal of Marketing Research, 16, (February), 64-73.

_____ and J. Paul Peter (1984), "Research Design Effects on the Reliability of Rating Scales: A Meta-Analysis," Journal of Marketing Research, 21 (November), 360-375.

Cooil, Bruce and Roland T. Rust (1994), "Reliability and Expected Loss: A Unifying Principle," Psychometrika, 59 (June), 203-216.

_____ and _____ (1995), "General Estimators for the Reliability of Qualitative Data," Psychometrika, 60 (June), 199-220.

Cronbach, Lee J., Nageswari Rajaratnam and Goldine C. Gleser (1963), "Theory of Generalizability: A Liberalization of Reliability Theory," British Journal of Statistical Psychology, 16 (November), 137-163.

_____, Goldine C. Gleser, Harinder Nanda, and Nageswari Rajaratnam (1972) The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles. New York: John Wiley & Sons.

Cronin, J. Joseph and Steven A. Taylor (1992), "Measuring Service Quality: A Reexamination and Extension," Journal of Marketing, 56 (July), 55-68.

_____ and _____ (1994), "SERVPERF Versus SERVQUAL: Reconciling Performance-Based and Perceptions-Minus-Expectations Measurement of Service Quality," Journal of Marketing, 58 (January), 125-131.

Devlin, Susan J. and H. K. Dong (1994), "Service Quality from the Customers Perspective," Marketing Research, 6 (1), 5-13.

_____, _____, and Marbue Brown (1993), "Selecting a Scale for Measuring Quality," Marketing Research, 5 (Summer), 12-17.

Farley, John U. and Donald R. Lehmann (1986), Meta-Analysis in Marketing: Generalization of Response Models. Lexington, MA: Lexington Books.

Fornell, Claes (1992), "A National Customer Satisfaction Barometer," Journal of Marketing, 56 (January), 6-21.

Gerbing, David W. and James C. Anderson (1988), "An Updated Paradigm for Scale Development Incorporating Unidimensionality and Its Assessment," Journal of Marketing Research, 25 (May), 186-192.

Hartley, H. O. and J. N. K. Rao (1967), "Maximum Likelihood Estimation for Mixed Analysis of Variance," Biometrika, 54, 93-108.

_____, J. N. K. Rao, and L. Lamotte (1978), "A Simple Synthesis-Based Method of Variance Component Estimation," Biometrics, 34, 233-244.

Hauser, John R., Duncan I. Simester, and Birger Wernerfelt (1994), "Customer Satisfaction Incentives," Marketing Science, 13 (Fall), 327-350.

Hughes, Marie Adele and Dennis E. Garrett (1988), "Inter-coder Reliability Estimation Approaches in Marketing: A Generalizability Theory Framework for Quantitative Data," Journal of Marketing Research, 27 (May), 185-95.

Iacobucci, Dawn, Kent A. Grayson, and Amy L. Ostrom (1994), "The Calculus of Service Quality and Customer Satisfaction: Theoretical and Empirical Differentiation and Integration," Advances in Services Marketing and Management, Vol. 3, 1-67.

Jacoby, Jacob (1978), "Consumer Research: A State of the Art Review," Journal of Marketing, 42 (April), 87-96.

Khuri, A. I. and Hardeo Sahai (1985), "Variance Components Analysis: A Selective Literature Survey," International Statistical Review, 53, 3, 279-300.

Lastovicka, John L. (1983), "Convergent and Discriminant Validity of Television Commercial Rating Scales," Journal of Advertising, 12 (2), 14-23.

Marcoulides, George A. (1995), "Designing Measurement Studies Under Budget Constraints: Controlling Error of Measurement and Power," Educational and Psychological Measurement, 55 (3), 423-428.

Narver, John C. and Stanley F. Slater (1990), "The Effect of A Market Orientation on Business Profitability," Journal of Marketing, 54 (4), 20-35.

Nunnally, Jum C. (1978), Psychometric Theory. Second Edition. New York: McGraw-Hill.

Parasuraman, A., Leonard L. Berry, and Valarie A. Zeithaml (1991), "Refinement and Reassessment of the SERVQUAL Scale," Journal of Retailing, 67 (Winter), 420-50.

_____, _____, and _____ (1993), "Research Note: More on Improving Service Quality Measurement," Journal of Retailing, 69 (Spring), 140-147.

_____, Valarie A. Zeithaml, and Leonard L. Berry (1985), "A Conceptual Model of Service Quality and its Implications for Future Research," Journal of Marketing, 49 (Fall), 41-50.

_____, _____, and _____ (1988), "SERVQUAL: A Multiple-Item Scale for Measuring Consumer Perceptions of Service Quality," Journal of Retailing, 64 (Spring), 12-40.

_____, _____, and _____ (1994a), "Reassessment of Expectations as a Comparison Standard in Measuring Service Quality: Implications for Further Research," Journal of Marketing, 58 (January), 111-124.

_____, _____, and _____ (1994b), "Alternative Scales for Measuring Service Quality: A Comparative Assessment Based on Psychometric and Diagnostic Criteria," Journal of Retailing, 70 (January), 201-230.

Peter, J. Paul (1979), "Reliability: A Review of Psychometric Basics and Recent Marketing Practices," Journal of Marketing Research, 16 (February), 6-17.

_____ and Michael L. Ray (1984), "Introduction," in Peter, J. Paul and Michael L. Ray (eds.) Measurement Readings for Marketing Research. Chicago, American Marketing Association, vii.

Peterson, Robert A. (1994), "A Meta-Analysis of Cronbach's Coefficient Alpha," Journal of Consumer Research, 21 (September), 381-391.

Rao, C. R. (1971), "Estimation of Variance and Covariance Components - MINQUE Theory," Journal of Multivariate Analysis, 1, 257-275.

Rentz, Joseph O. (1987), "Generalizability Theory: A Comprehensive Method for Assessing and Improving the Dependability of Marketing Measures," Journal of Marketing Research, 24 (February), 19-28.

_____ (1988), "An Exploratory Study of the Generalizability of Selected Marketing Measures," Journal of the Academy of Marketing Science, 16 (Spring), 141-150.

Reibstein, David, John E. Bateson and William Boulding (1988), "Conjoint Analysis Reliability: Empirical Findings," Marketing Science, 7 (Summer), 271-286.

Rust, Roland T. and Bruce Cooil (1994), "Reliability Measures for Qualitative Data: Theory and Implications," Journal of Marketing Research, 31 (February), 1-14.

_____, Anthony J. Zahorik, and Timothy L. Keiningham (1995), "Return on Quality (ROQ): Making Service Quality Financially Accountable," Journal of Marketing, 59 (April), 58-70

Salkin, H. M. (1975), Integer Programming. Reading, MA: Addison-Wesley.

Sanders, P. F., T. J. J. M. Theunissen, and S. M. Baas (1989), "Minimizing the Number of Observations: A Generalization of the Spearman-Brown Formula," Psychometrika, 54, 587-598.

_____, _____, and _____ (1991), "Maximizing the Coefficient of Generalizability under the Constraint of Limited Resources," Psychometrika, 56, 87-96.

Satterthwaite, F. E. (1946), "An Approximate Distribution of Estimates of Variance Components," Biometrics Bulletin, Vol. 2, 110-114.

Searle, Shayle R., George Casella, and Charles E. McCulloch (1992), Variance Components. New York: John Wiley and Sons.

Schmalensee, Diane H. (1994), "Finding the Perfect Scale," Marketing Research, 6 (Fall) 24-27.

Shavelson, Richard J. and Noreen M. Webb (1991), Generalizability Theory: A Primer. Newbury Park, CA: SAGE Publications.

Spechler, Jay W. (1991) When America Does It Right: Case Studies in Service Quality. Norcross, GA: Institute of Industrial Engineers.

Steenkamp, Jan-Benedict E. M. and Hans C. M. van Trijp (1991), "The Use of LISREL in Validating Marketing Constructs," International Journal of Research in Marketing, 8 (November), 283-299.

Teas, Kenneth R. (1994), "Expectations as a Comparison Standard in Measuring Service Quality: An Assessment of a Reassessment," Journal of Marketing, 58 (January), 132-139.

Zahorik, Anthony J. and Roland T. Rust (1992), "Modeling the Impact of Service Quality on Profitability: A Review," Advances in Services Marketing and Management, Vol. 1, 247-276.

Chapter 3

Influence of Estimation Method and Interval Estimates on Optimality of Measurement Designs

A methodology to optimize the design of future measurement studies was a major contribution of the paper by Finn and Kayandé (1997). They also suggested a general framework, based on generalizability theory (Cronbach et al. 1972), to assess the generalizability of measurement, which they illustrated with a study on service quality measurement. Variance component models, commonly known as mixed effect models, are the engine which allowed the framework to be put into practice. There is a large literature in statistics on the point and interval estimation of such models (Khuri and Sahai 1985, Searle, Casella, and McCulloch 1992, Burdick and Graybill 1992). This literature suggests that the estimates obtained from different estimation methods can vary with factors such as the extent of sampling on each effect and whether the collected data make up a complete balanced design. In addition, the estimation methods are built around different assumptions. If the estimates of variance components depend on the method of estimation, it is obvious that the "optimal" designs generated by Finn and Kayandé (1997) may also depend on the method of estimation. In addition, Shavelson and Webb (1981) suggest that the variability of variance component estimates may prove to be the *achilles' heel* of generalizability theory-based approaches to measurement. In statistical terms, this simply implies that the interval estimates of variance components might be too wide to make any definitive conclusions about either optimal designs or generalizability of any specific design.

In this paper, we review different methods to obtain point and interval estimates of variance components. In addition, we use the data collected by Finn and Kayandé (1997) to empirically illustrate the statistical estimation issues raised in the review. The paper is structured as follows. We first provide a brief introduction to the different estimation methods, specially in the context of balanced and unbalanced data/designs. Then, we

present the general model underlying variance component models. Subsequent to the presentation of the general model, we present a detailed explanation of the statistical theory underlying each method of point estimation, and the associated issue of sampling variability of the estimates, independent of the "balance" in the design/data. Throughout this discussion of methods, we offer comments on the appropriateness of each method in the presence of unbalanced data/designs. We then discuss methods to obtain interval estimates, suggesting the appropriateness of each method in the presence of unbalanced data/design. In the empirical section, we explore whether the estimates of variance components depend on methods of estimation, determine the impact of the estimation method on the optimal designs for measurement studies in the context of Finn and Kayandé's (1997) data, and investigate whether interval estimates might put into question the practical usefulness of framework developed by Finn and Kayandé (1997). Finally, drawing upon the empirical results and the theoretical section, we present recommendations for future research on measurement. The recommendations range from the type of estimation method that is most useful for a given set of data to the importance of estimating confidence intervals in measurement research.

3.1 Methods for Estimating Variance Components

The ability of estimation methods to produce variance component estimates of desirable properties is dependent largely on the balance in the data or design. Unbalanced data are of two types, both of which are particularly likely to arise in marketing studies. First, unbalanced data can occur because of missing observations. Observations are missing because either some respondents lack experience with some levels of objects or some respondents overlook an item in a questionnaire. Second, the design of the study could be unbalanced, i.e., the number of levels within a nested factor could be unequal. Both types of unbalanced data create difficulties when trying to use methods such as analysis of variance for estimating variance components. We now give a brief description

of several methods of estimation, with the attendant advantages and disadvantages of each method, specially in the context of the balance in the data.

The traditional analysis of variance (ANOVA) method of estimation is the most intuitive, because of the direct relationship of the estimates to the mean squares, sums of squares, and degrees of freedom. However, the method produces estimates of variance with an unknown distribution, and therefore confidence intervals are difficult to construct (although we recognize that bootstrapping can be used to obtain confidence intervals when the theoretical distribution is not known, we restrict the discussion in this paper to theoretical methods). Additionally, the ANOVA method should be avoided in the presence of unbalanced data, because the estimates with such data are biased, not unique, and methods are essentially ad hoc (Blischke 1968, Searle, Casella, and McCulloch 1992, Rao 1971b). The potential for negative variance estimates is another serious problem with ANOVA methods.

The weaknesses of the estimates produced by ANOVA led to the search for alternative methods. The computational difficulties associated with maximum likelihood methods severely limited their use until a landmark paper was published by Hartley and Rao (1967). Advances in computational ability have now allowed these methods to become popular with statisticians. Maximum likelihood methods require attributing a distribution to the data, and most closed form solutions have been derived using an assumption of normality. The estimates produced by maximum likelihood (ML) are optimal with very well known distributional properties, even in the presence of unbalanced data. Additionally, maximum likelihood methods preclude the possibility of negative estimates; although an advantage, this may result in the occurrence of biased estimates. Maximum likelihood methods include restricted maximum likelihood (REML), which is useful for models that include fixed effects. The REML method maximizes that part of the likelihood function that is invariant to the fixed effects. The traditional ML method allows the estimation of fixed effects simultaneously with

variance components, which is not possible with REML. The major disadvantage of the maximum likelihood method is that it is a large sample method and properties of estimates are based on asymptotic normal theory. The sensitivity of these estimates to the violation of large sample assumptions is of concern, primarily because of the real possibility of few levels to factors in a study. Measurement studies in marketing have typically employed few levels of all factors other than respondents. On the other hand, the major advantages are the optimal properties of estimates even in the presence of unbalanced data and the absence of negative variance estimates.

Lamotte (1970, 1971) and Rao (1970, 1971a) describe a method that is not as computationally burdensome as ML methods because of its non-iterative nature. They developed the minimum norm quadratic unbiased (MINQUE) estimation method, by requiring that the quadratic unbiased estimator minimize a Euclidean norm. This Euclidean norm becomes the variance in the presence of normality. An important aspect of the method is the assignment of *a priori* values for the variance components. The method does not require normality, although the assumption of normality makes the interpretation easier. The advantage of the method is that it does not solve the equations iteratively as do the maximum likelihood procedures. This advantage results in computationally less burdensome procedures. However, the assignment of *a priori* values to the variance components is also a disadvantage because of the potential sensitivity of the results to differences in the pre-assigned values. For example, two different researchers can get different results from the same data set by using different *a priori* values (Searle, Casella, and McCulloch 1992). This problem has been resolved by Hartley, Rao and Lamotte (1978) by pre-assigning zeros for all variance components (except the error variance component, which is set to 1), thus avoiding the problem of non-unique estimators to some extent. However, the estimates can potentially be negative and are often biased.

In addition to the ANOVA, maximum likelihood, and MINQUE methods, Bayesian methods can also be used for estimation of variance components. In this paper, we focus on only classical frequentist approaches to variance component estimation. There are two reasons for this focus. First, classical approaches have been tested and compared with each other more than has been the case for Bayesian methods. Second, several simulation studies suggest that maximum likelihood methods are more efficient than Bayesian or classical ANOVA methods for variance components estimation (Khuri and Sahai 1985), specially in the presence of unbalanced data. We now present the statistical model underlying all variance component estimation methods.

3.1.1 General Statistical Model

The most general form of a linear model is a mixed model with both fixed and random effects. With no fixed effects, the model reduces to a random effects model. The distinction between fixed and random effects is essentially based on the sampling assumptions for each factor. If the levels of a factor have been randomly sampled, the effects associated with the factor are treated as random. If the levels of a factors have been selected specifically because the researcher is interested in *those levels*, the effect associated with the factor is treated as fixed. We refer the reader to Searle (1987), Searle, Casella, and McCulloch (1992), and Shavelson and Webb (1981) for details on the distinction between fixed and random effects. In this paper, all effects have been treated as random. However, for the sake of generality, we present the more general mixed model, with fixed and random effects. Consideration of the random effects model as a special case of the mixed effects model is straightforward.

The general mixed model can be expressed as follows (Searle 1987),

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e} \tag{3.1}$$

where $\mathbf{y}$ is a vector of observations, $\mathbf{X}$ is the design matrix associated with fixed effects, β is a vector of fixed effects associated with design matrix $\mathbf{X}$, $\mathbf{u}$ is the vector of random

effects corresponding to the design matrix $\mathbf{Z}$ for random effects, and $\mathbf{e}$ is a vector of error terms defined as $\mathbf{e} = \mathbf{y} - E(\mathbf{y} \mid \mathbf{u})$, and

(3.2) $$E(\mathbf{y} \mid \mathbf{u}) = \mathbf{X}\beta + \mathbf{Zu}$$

(3.3) $$E(\mathbf{y}) = \mathbf{X}\beta$$

Thus, if there are no fixed effects in the model the expected value of $\mathbf{y}$ is $\mathbf{0}$.

From Equation 3.1, the variance-covariance matrix $\mathbf{V}$ of y can be defined as,

(3.4) $$\mathbf{V} = \text{var}(\mathbf{Zu} + \mathbf{e})$$

Assuming that $\text{cov}(\mathbf{u}, \mathbf{e}') = \mathbf{0}$ and $\text{var}(\mathbf{e}) = \sigma_e^2 \mathbf{I}$ allows for an expression for $\mathbf{V}$ as,

(3.5) $$\mathbf{V} = \mathbf{Z}\,\text{var}(\mathbf{u})\mathbf{Z}' + \sigma_e^2 \mathbf{I}$$

The vector u can be partitioned into r sub-vectors,

(3.6) $$\mathbf{u}' = [\mathbf{u}_1' \;\; \mathbf{u}_2' \;\; \ldots \;\; \mathbf{u}_i' \;\; \ldots \;\; \mathbf{u}_r']$$

where each sub-vector $\mathbf{u}_i$ has as elements the effects corresponding to all levels of the i^{th} random effect. Similarly $\mathbf{Z}$ is partitioned as,

(3.7) $$\mathbf{Z} = [\mathbf{Z}_1 \;\; \mathbf{Z}_2 \;\; \ldots \;\; \mathbf{Z}_i \;\; \ldots \;\; \mathbf{Z}_r]$$

This results in writing the model in Equation 3.1 as,

(3.8) $$\mathbf{y} = \mathbf{X}\beta + \sum_{i=1}^{r} \mathbf{Z}_i\mathbf{u}_i + \mathbf{e}$$

and therefore,

(3.9) $$\mathbf{V} = \text{var}\left(\sum_{i=1}^{r} \mathbf{Z}_i\mathbf{u}_i + \mathbf{e}\right)$$

Now we make assumptions for the sub-vectors $\mathbf{u}_i$,

(3.10) $E(\mathbf{u}_i) = \mathbf{0}$, $\forall$ i,

(3.11) $\text{var}(\mathbf{u}_i) = \sigma_i^2 \mathbf{I}_{q_i}$, $\forall$ i and where q_i is the number of levels of $\mathbf{u}_i$,

(3.12) $\text{cov}(\mathbf{u}_i, \mathbf{u}'_{i'}) = \mathbf{0}$, $\forall\, i \neq i'$, and

(3.13) $\text{cov}(\mathbf{u}_i, \mathbf{e}') = \mathbf{0}$, $\forall$ i.

Using these assumptions in Equations 3.10 through to 3.13, Equation 3.9 can be written as,

(3.14) $$\mathbf{V} = \sum_{i=1}^{r} \mathbf{Z}_i \mathbf{Z}_i' \sigma_i^2 + \sigma_e^2 \mathbf{I}$$

Defining further that $\mathbf{u}_o = \mathbf{e}$, $\sigma_0^2 = \sigma_e^2$, and $\mathbf{Z}_o = \mathbf{I}$, the most general and common form of the mixed model can be expressed as,

(3.15) $$\mathbf{y} = \mathbf{X}\beta + \sum_{i=0}^{r} \mathbf{Z}_i \mathbf{u}_i$$

The dispersion matrix of **y** can then be written as *the general variance components model* (Hartley and Rao 1967),

(3.16) $$\mathbf{V} = \sum_{i=0}^{r} \mathbf{Z}_i \mathbf{Z}_i' \sigma_i^2$$

The components of variance can be estimated by different methods, which are now discussed in detail.

3.1.2 Analysis of Variance Estimation

The general ANOVA method of estimation is based on expressing expected mean squares as a linear combination of variance components and then solving for the variance components. The method is computationally and intuitively appealing, at least in the case of balanced data. The familiarity of ANOVA procedures allows researchers to interpret the variance components with greater ease.

The general method involves equating the mean squares from an ANOVA to the expected values of the mean squares. The expected values of mean squares can be expressed as linear combinations of the variance components, which are estimated by solving the system of equations involving these linear combinations. Let **m** be the vector of mean squares, with the same order as σ^2, the vector of variance components. Then, there is a **P** such that,

(3.17) $$E(\mathbf{m}) = \mathbf{P}\sigma^2$$

Then the ANOVA estimator of σ^2 is $\hat{\sigma}^2$, obtained from,

(3.18) $$\hat{\sigma}^2 = \mathbf{P}^{-1}\mathbf{m}, \text{ with } \mathbf{P} \text{ nonsingular.}$$

These ANOVA estimators are unbiased in the case of balanced data. Graybill and Hultquist (1961) also established that these estimators have minimum variance of all estimators that are quadratic functions of the observations and unbiased. In the case of balanced data and making an assumption of normality, these estimators are also minimum variance of all unbiased estimators (Graybill 1954, Graybill and Wortham 1956).

The sampling variances of the estimators can be derived because the variance components are linear functions of χ^2 variables (expected mean squares). However, in general, there is no closed form expression for the distribution of most estimators and the variances of the estimators are functions of unknown components.

From Equation 3.17, the variance in $\hat{\sigma}^2$ can be written as,

$$\text{var}(\hat{\sigma}^2) = \mathbf{P}^{-1}\,\text{var}(\mathbf{m})\,\mathbf{P}^{-1'} \tag{3.19}$$

The variance of a mean square can be expressed as,

$$\text{var}(M_i) = \frac{2\left[E(M_i)\right]^2}{f_i}, \tag{3.20}$$

where f_i are the degrees of freedom associated with the mean square M_i.

It can therefore be shown that an unbiased estimator of the variance of $\hat{\sigma}^2$ is,

$$\text{var}(\hat{\sigma}^2) = \mathbf{P}^{-1}\left\{\frac{2M_i^2}{f_i+2}\right\}_d\mathbf{P}^{-1'} \tag{3.21}$$

where the notation "$\{\}_d$" indicates that the term in the bracket is a diagonal matrix.

This estimate of the sampling variance of the variance component estimates can only serve as an indication of the dispersion of the estimates. Little can be done, however, in terms of constructing confidence intervals because of the inability to derive the exact or approximate form of the distribution of the estimates. To construct intervals, we use methods suggested by Burdick and Graybill (1992), which have been shown to work well for samples of all sizes. The methods are presented later in the paper in the section on confidence interval estimation.

3.1.3 Maximum Likelihood Estimation

Maximum likelihood estimation is tractable with an assumption of normality of the random effects and the error terms. Assuming normality, the model in Equation 3.1 can be written as,

(3.22) $$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e} \sim N(\mathbf{X}\beta, \mathbf{V})$$

with $\mathbf{V} = \sum_{i=0}^{r} \mathbf{Z}_i\mathbf{Z}_i' \sigma_i^2$

The likelihood function is then,

(3.23) $$L = (2\pi)^{-N/2} |\mathbf{V}|^{-1/2} \exp[(-1/2)(\mathbf{y} - \mathbf{X}\beta)' \mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\beta)]$$

The likelihood function L is converted to the log-likelihood function l, then maximized by taking the derivatives of l with respect to β and σ_i^2, and setting the resulting equations to zero. The derivation of the equations is given in Searle, Casella, and McCulloch (1992, pg. 235-236). The resulting maximum likelihood solutions are,

(3.24) $$\mathbf{X}'\tilde{\mathbf{V}}^{-1}\mathbf{X}\beta = \mathbf{X}'\tilde{\mathbf{V}}^{-1}\mathbf{y}$$

(3.25) $$\mathrm{tr}\left(\tilde{\mathbf{V}}^{-1}\mathbf{Z}_i\mathbf{Z}_i'\right) = (\mathbf{y} - \mathbf{X}\beta)'\tilde{\mathbf{V}}^{-1}\mathbf{Z}_i\mathbf{Z}_i'\tilde{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\beta), \text{ for } i, j = 0, 1, 2, 3, \ldots, r.$$

These equations are then solved for β and σ^2. In addition, the second derivatives of l with respect to β and σ_i^2 are examined along with the parameter space restrictions, which in case of maximum likelihood are the non-negativity of all variance components except the error variance component, which is constrained to be positive. The maximum likelihood solution is estimated iteratively using this specification. The advantage is that the fixed effects β are simultaneously estimated along with the variance components σ^2.

Searle (1987) and Searle, Casella, and McCulloch (1992) provide an algebraically simpler expression for the equations by defining,

(3.26) $$\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}$$

and

(3.27) $$\mathbf{I} = \mathbf{V}^{-1}\mathbf{V} = \mathbf{V}^{-1}\sum_{i=0}^{r} \mathbf{Z}_i\mathbf{Z}_i' \sigma_i^2$$

Using equations 3.25 and 3.26 together with Equation 3.24 provides the following formulation for the equations to estimate σ^2:

$$\left\{\mathrm{tr}\left(\tilde{\mathbf{V}}^{-1}\mathbf{Z}_i\mathbf{Z}_i'\tilde{\mathbf{V}}^{-1}\mathbf{Z}_j\mathbf{Z}_j'\right)\right\}\left\{\sigma_i^2\right\} = \left\{\mathbf{y}'\tilde{\mathbf{P}}\mathbf{Z}_i\mathbf{Z}_i'\tilde{\mathbf{P}}\mathbf{y}\right\} \tag{3.28}$$

for i, j = 0, 1, 2, 3, . . . , r.

The estimates of the dispersion matrix $\Sigma_{\hat{\sigma}^2}$ of the estimates of the variance components, in the case of maximum likelihood, are given by,

$$\Sigma_{\hat{\sigma}^2} = 2\left[\left\{{}_m\mathrm{tr}\left(\tilde{\mathbf{V}}^{-1}\mathbf{Z}_i\mathbf{Z}_i'\tilde{\mathbf{V}}^{-1}\mathbf{Z}_j\mathbf{Z}_j'\right)\right\}_{i,j=0}^{r}\right]^{-1} \tag{3.29}$$

where the notation "${}_m$" indicates that the term in the bracket is a matrix.

3.1.4 Restricted Maximum Likelihood Estimation

Patterson and Thompson (1971) developed the restricted maximum likelihood method for general mixed models in order to account for fixed effects in the maximum likelihood estimation. Thus REML is a variant of the ML method, where the estimators are obtained by maximizing only that part of the likelihood that is invariant to the location parameter in the model in Equation 3.1, i.e., the fixed effects $\mathbf{X}\beta$. The procedure is given in detail in Searle, Casella, and McCulloch (1992, pg. 249-251). The resulting (r + 1) equations to be set to zero to estimate the variance components are,

$$\left\{\mathrm{tr}\left(\tilde{\mathbf{P}}\mathbf{Z}_i\mathbf{Z}_i'\tilde{\mathbf{P}}\mathbf{Z}_j\mathbf{Z}_j'\right)\right\}\left\{\sigma_i^2\right\} = \left\{\mathbf{y}'\tilde{\mathbf{P}}\mathbf{Z}_i\mathbf{Z}_i'\tilde{\mathbf{P}}\mathbf{y}\right\} \tag{3.30}$$

for i, j = 0, 1, 2, 3, . . . , r.

The resulting (r + 1) equations are identical to the ML Equations 3.28, except that $\tilde{\mathbf{V}}^{-1}$ is replaced by $\tilde{\mathbf{P}}$.

REML also has the computational burdens of the maximum likelihood estimation. However, the REML estimates are *unbiased* in the case of balanced data, unlike the ML estimates. Also, in the case of balanced data, the solutions to the REML Equations 3.30 are identical to the ANOVA estimates, whether or not normality is assumed, and are therefore minimum variance unbiased. However, note that these solutions are not REML

estimators *unless* the two conditions of normality *and* non-negativity of estimates are satisfied. Thus, when there are negative estimates, the ANOVA and REML solutions should not be expected to be similar. In the case of unbalanced data, both REML and ML estimators are biased. The REML method does not provide direct estimates of the fixed effects in the mixed model, which are provided by the ML method.

The estimates of the dispersion matrix $\sum_{\hat{\sigma}^2}$ of the estimates of the variance components, in the case of restricted maximum likelihood, are given by,

$$\sum_{\hat{\sigma}^2} \cong 2\left[\left\{{}_{m}\mathrm{tr}\left(\mathbf{PZ}_i\mathbf{Z}_i'\mathbf{PZ}_j\mathbf{Z}_j'\right)\right\}_{i,j=0}^{r}\right]^{-1} \tag{3.31}$$

3.1.5 MINQUE (Minimum Norm Quadratic Unbiased Estimation)

The ANOVA method of estimation yields estimators with unknown properties for the most part. This prompted Lamotte (1971, 1972) and Rao (1971a, b, 1972) to develop a method of estimation that would attribute desirable properties to the estimates. The maximum likelihood methods lead to estimates being consistent, efficient, and asymptotically normal. The important difference between ML and ANOVA methods on the one hand and MINQUE on the other, is that the former methods result in estimates with certain properties, whereas the latter method specifies these properties at the outset. Therefore, MINQUE method is called a criteria based procedure. There are several methods that fall under this category. We describe the general MINQUE procedure and the MIVQUE(0) procedure, the latter method being used for the estimation of variance components in the empirical illustration.

The method suggested by Rao (1970, 1971a, b, 1972) does not require the assumption of normality and leads to estimates that have minimum variance of all unbiased estimates that are quadratic functions of the observations. Searle, Casella, and McCulloch (1992) provide a derivation of the results produced in this section. Let **w** represent the set of weights that are the *a priori* values chosen by a researcher for the variance components. The vector **w** is of the same order as σ^2. The matrix **V** is now

denoted by $\mathbf{V}_w$, with the weights **w** in place of σ^2. Similarly, denote **P** of Equation 3.26 by $\mathbf{P}_w$, with $\mathbf{V}_w$ replacing **V**. Then the estimation equations for MINQUE are similar to the ML equations 3.28 and are given by,

$$\left\{\mathrm{tr}\left(\mathbf{P}_w\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}_w\mathbf{Z}_j\mathbf{Z}_j'\right)\right\}\left\{\sigma_i^2\right\} = \left\{\mathbf{y}'\mathbf{P}_w\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}_w\mathbf{y}\right\} \tag{3.32}$$

for i, j = 0, 1, 2, 3, . . . , r.

A modification of the MINQUE method is the minimum variance quadratic estimation method (MIVQUE(0)). The modification assumes normality, and additionally assumes *a priori* values of 0 for all variance components with exception of σ_e^2, which is assumed to be 1. The MIVQUE(0) method produces estimates that are minimum variance quadratic unbiased.

An important assumption made in the estimation by MINQUE methods is that the *a priori* values assigned to variance components are the true values of the variance components. Only then are the estimates locally minimum variance quadratic unbiased (Sahai and Khuri 1985). This limitation is a result of the non-iterative nature of MINQUE methods.

3.2 Methods to Determine Interval Estimates of Variance Components

Burdick and Graybill (1992) and Wonnacott (1987) argue for the use of confidence intervals in variance component modelling because interval estimates provide the "most informative summary" results of statistical inference. In addition, interval estimates also provide an indication of the confidence we may be able to place in the ability of an "optimal" measurement design, derived from point estimates, to provide information of acceptable levels of generalizability.

Following this argument, we present a selection of the general methods available for constructing confidence intervals around variance component estimates. The dispersion matrices from ML estimation can be used to derive confidence intervals. Such confidence intervals are easy to derive because of the assumption of normality. In the

empirical section, we present confidence intervals on estimates derived from ML methods. The problem with using the dispersion matrices from ML estimation for constructing confidence intervals is that the ML methodology is based on large-sample asymptotic normal theory. Large samples are rarely possible for most factors in marketing surveys, and have rarely been used in most studies which use variance component modelling. The alternative is to use approximate methods that work for all sample sizes, small or large, although they are based on normal theory and require the use of ANOVA estimators. Thus, the problem of using ANOVA estimators with unbalanced data creates a trade-off between using the asymptotically normal large-sample dispersion matrix of maximum likelihood estimators and the methods suggested by Burdick and Graybill (1992) for samples of all sizes. Burdick and Graybill (1992) review a large number of methods to construct confidence intervals on estimates obtained from ANOVA methods. We present two methods in this paper; we refer the reader to their book to obtain details of other methods.

3.2.1 Satterthwaite's Method:

The history of interval estimation for variance components requires mention of the methodology developed by Satterthwaite (1941, 1946). This method continues to be popular with practitioners because of its ease of use and intuitive appeal (Burdick and Graybill 1992, Brennan 1983). Approximate intervals for estimates of confidence intervals can be constructed with the Satterthwaite method for almost any experimental design. The general methodology is to express any variance component as a linear combination of the expected mean squares from an analysis of variance, and then attribute an approximate distribution to this combination. Any variance component in a variance component model can be expressed as E(MS′) – E(MS″), where,

(3.33) $$MS' = \sum_{i=1}^{n} f_i MS_i$$

(3.34) $$MS'' = \sum_{i=n+1}^{N} f_i MS_i$$

where N is the number of mean squares that form the expression for the variance component in question, f_i is the coefficient associated with the mean square MS_i. An approximate F-test for the significance of the estimates is given by,

(3.35) $$F = \frac{MS'}{MS''} \sim F_{p,q}$$

where,

(3.36) $$p = \frac{\left(\sum_{i=1}^{n} f_i MS_i\right)^2}{\sum_{i=1}^{n} \left(f_i MS_i\right)^2 / df_i},$$

(3.37) $$q = \frac{\left(\sum_{i=n+1}^{N} f_i MS_i\right)^2}{\sum_{i=n+1}^{N} \left(f_i MS_i\right)^2 / df_i},$$

and df_i is the number of degrees of freedom associated with the mean square MS_i.

An approximate 100(p)% interval of an estimated variance component is then given by,

(3.38) $$Prob\left[\frac{\sigma_i^2 \nu}{\chi_U^2(\nu)} \le \sigma_i^2 \le \frac{\sigma_i^2 \nu}{\chi_L^2(\nu)}\right] \approx p$$

where $\nu = \dfrac{\left(\sum_i f_i MS_i\right)^2}{\sum_i \left(f_i MS_i\right)^2 / df_i}$ are the effective degrees of freedom; $\chi_U^2(\nu)$ and $\chi_L^2(\nu)$ are the lower U = (1 + p)/2 and L = (1 – p)/2 percentage points of the chi-squared distribution with ν degrees of freedom.

The effective degrees of freedom is not generally an integer, and is thus rounded off to the greatest integer less than or greater than ν. Satterthwaite's procedure is recommended when the degrees of freedom df_i are all large or, when small, all equal.

When there are large differences in the degrees of freedom, the procedure produces unacceptably liberal[1] confidence intervals (Burdick and Graybill 1992). Also, if some of the f_i's are positive and some negative, Burdick and Graybill (1992) strongly recommend against the procedure. Additionally, Khuri and Sahai (1985) strongly question the logic of using this procedure with unbalanced data because the mean squares are no longer independent and do not have a distribution of a scalar multiple of a independent chi-squared variables. Thus, the procedure is suitable for only a specific class of problems, all of which require balanced data. Finn and Kayandé (1997) used Satterthwaite's method on the balanced subset of their data, to show that most of the non-negative estimates of variance components in their study were significantly different from zero.

Several other papers (e.g., Welch 1956, Graybill and Wang 1980) have recommended approximate large-sample confidence intervals that suffer from similar limitations. Although Graybill and Wang's (1980) procedure works best of all these methods in terms of producing a confidence coefficient close to the true confidence coefficient, all of these methods do not work well in the presence of small samples (implying low number of levels for a factor in a study).

We now present the general procedure suggested by Ting, Burdick, Graybill, Jeyaratnam, and Lu (1990), which overcomes the major limitations of these methods.

3.2.2 General Procedure to Construct Confidence Intervals with ANOVA estimates:

A variance component can be expressed as a linear combination of expected mean squares. The coefficients of mean squares in these linear combinations are commonly positive *and* negative. Satterthwaite's method should not be used with negative coefficients and hence Ting et al. (1990) suggested a general procedure that can be used with negative and positive coefficients in the linear combination of mean squares. In several simulation studies, Ting et al.(1990) show that their method, although cumbersome, produces confidence coefficients close to the true confidence coefficient even in the context of small samples.

The notation for this procedure differs from that of Satterthwaite's method and is therefore explained in some detail. The estimate $\hat{\delta}$ of a variance component can be expressed as,

(3.39) $$\hat{\delta} = \sum_{q=1}^{P} c_q S_q^2 - \sum_{r=P+1}^{Q} c_r S_r^2$$

where, S_i^2's are the mean squares and c_i's are the coefficients associated with them.

The lower bound for an upper $(1 - \alpha)$ interval on δ is given by,

(3.40) $$L = \hat{\delta} - \sqrt{V_L}$$

where

$$V_L = \sum_{q=1}^{P} G_q^2 c_q^2 S_q^4 + \sum_{r=P+1}^{Q} H_r^2 c_r^2 S_r^4 + \sum_{q=1}^{P} \sum_{r=P+1}^{Q} G_{qr} c_q c_r S_q^2 S_r^2 + \sum_{q=1}^{P-1} \sum_{t>q}^{P} G_{qt}^* c_q c_t S_q^2 S_t^2$$

$$G_q = 1 - \frac{1}{F_{\alpha:\, n_q, \infty}} \qquad (q = 1, \ldots, P)$$

$$H_r = \frac{1}{F_{1-\alpha:\, n_r, \infty}} - 1 \qquad (r = P + 1, \ldots, Q)$$

$$G_{qr} = \frac{(F_{\alpha:\, n_q, n_r} - 1)^2 - G_q^2 F_{\alpha:\, n_q, n_r}^2 - H_r^2}{F_{\alpha:\, n_q, n_r}}$$

$$G_{qt}^* = \left(\frac{1}{P-1}\right)\left[\left(1 - \frac{1}{F_{\alpha:\, n_q + n_t, \infty}}\right)^2 \frac{(n_q + n_t)^2 - H_r^2}{n_q n_t} - \frac{G_q^2 n_q}{n_t} - \frac{G_t^2 n_t}{n_q}\right] \qquad (\text{for } t = q + 1, \ldots, P)$$

Similarly, the upper bound on a lower $(1 - \alpha)$ interval on δ is given by,

(3.41) $$U = \hat{\delta} + \sqrt{V_U}$$

where,

$$V_U = \sum_{q=1}^{P} H_q^2 c_q^2 S_q^4 + \sum_{r=P+1}^{Q} G_r^2 c_r^2 S_r^4 + \sum_{q=1}^{P} \sum_{r=P+1}^{Q} H_{qr} c_q c_r S_r^2 S_q^2 + \sum_{r=P+1}^{Q-1} \sum_{u>r}^{Q} H_{ru}^* c_r c_u S_r^2 S_u^2$$

$$H_q = \frac{1}{F_{1-\alpha:\, n_q, \infty}} - 1 \qquad (q = 1, \ldots, P)$$

$$G_r = 1 - \frac{1}{F_{\alpha:\, n_r, \infty}} - 1 \qquad (r = P + 1, \ldots, Q)$$

$$H_{qr} = \frac{(1 - F_{1-\alpha:\, n_q, n_r})^2 - H_q^2 F_{1-\alpha:\, n_q, n_r}^2 - G_r^2}{F_{1-\alpha:\, n_q, n_r}}$$

$$H_{ru}^* = \left(\frac{1}{Q - P - 1}\right)\left[\left(1 - \frac{1}{F_{\alpha:\, n_r + n_u, \infty}}\right)^2 \frac{(n_r + n_u)^2 - H_r^2}{n_r n_u} - \frac{G_r^2 n_r}{n_u} - \frac{G_u^2 n_u}{n_r}\right] \quad (\text{for } u = r + 1, \ldots, Q)$$

3.2.3 Confidence intervals on ratios of variance components:

The interval estimation of ratios of variance components has also been given considerable attention by researchers in statistics (Burdick and Graybill 1992). Such ratios are of interest because they represent the proportion of total variance accounted for by a factor, or a signal to noise ratio, or an intraclass correlation coefficient (a generalizability coefficient). Ratios are of interest in this paper, specifically in the context of the question raised early in this paper about the impact of interval estimates on the optimality of measurement design. The ability to produce interval estimates of variance components does not imply that we have an answer to the question. The reason is that it is not straightforward to use the interval estimates to derive an interval estimate for most ratios of variance components, or the generalizability coefficient. The estimate of the generalizability coefficient drives the optimization process, and therefore it is important to obtain an estimate of the sampling variability of the estimated coefficient.

The problem with estimating the confidence intervals around ratios of variance components lies in the inability to attribute exact distributions to the numerator or the denominator of most ratios. Burdick and Graybill (1992) offer approximate confidence intervals for some ratios of variance components estimated from simple designs and also inform us that there is no general method available to form interval estimates of ratios. For simple, 2 or 3 factor designs (nested or crossed), Burdick and Graybill (1992) provide

confidence intervals for ratios that result in confidence coefficients close to the true confidence coefficients. The procedures can be generalized to unbalanced data. However, in this paper, we do not explore this possibility because of the uncertain state of the literature regarding designs with more than three factors, specially with unbalanced data.

There is, however, one special case of ratios which deserves attention. In a generalizability theory approach to measurement, it is common to estimate a generalizability coefficient, *assuming a specific design* for a decision study to be conducted in the future. If the data from a generalizability study were to be used for the purposes of making decisions (instead of conducting another study for making decisions), then the generalizability coefficient should reflect the sample sizes used in the generalizability study. Schroeder and Hakstian (1990) use this possibility to develop a method to estimate approximate intervals on the generalizability coefficient. The procedure is only valid for balanced data, and only for a design that has sample sizes equal to those in the study used to estimate the variance components.

Schroeder and Hakstian's (1990) methodology to derive a confidence interval around a generalizability coefficient is as follows. A generalizability coefficient $E\hat{\rho}^2$, for a design with sample sizes equal to those used to estimate variance components, can be expressed as a function of the mean squares M_i (i = 1, 2, . . ., I) estimated in the study. Then, it can further shown that $(1-E\hat{\rho}^2)/(1-E\rho^2)$ follows an approximate F-distribution with degrees of freedom df_N and $(n_{object}-1)$. df_N is defined by Satterthwaite's (1941, 1946) method for a combination of mean squares (see section on Satterthwaite's method in this paper) and n_{object} are the number of levels of the object of measurement. The approximate F-distribution results from the mean squares in the numerator and denominator, which are independent, following chi-square distributions. Further, using Paulson's (1942) transformation, it can be shown that,

$$(3.42) \qquad (1-E\hat{\rho}^2)^{\frac{1}{3}} \sim NID\left[c\,(1-E\hat{\rho}^2)^{\frac{1}{3}}, \text{var}(1-E\hat{\rho}^2)^{\frac{1}{3}}\right]$$

Schroeder and Hakstian (1990) show that 'c' is a function of the degrees of freedom df_N and n_{object}, and can be ignored because there is a negligible loss in precision as a result. Using the delta method (Rao 1973, p. 387) to estimate sampling variance of an estimator, Schroeder and Hakstian (1990) further show that the variance of $\theta = (1 - E\hat{\rho}^2)^{1/3}$ can be expressed as,

$$\text{var}\,\theta = \Phi' \Sigma \Phi \tag{3.43}$$

where,

$$\Phi' = \left[\frac{\partial\theta}{\partial MS_1} \quad \frac{\partial\theta}{\partial MS_2} \quad \cdots \quad \frac{\partial\theta}{\partial MS_I} \right]$$

and each element of the diagonal matrix Σ is $\dfrac{2\left[E(M_i)\right]^2}{df_i}$, $(i = 1, 2, \ldots, I)$, and where df_i are the degrees of freedom associated with the mean square M_i. Thus, an approximate 90% confidence interval around the generalizability coefficient is given by,

$$\left[1 - \left((1 - E\hat{\rho}^2)^{\frac{1}{3}} \pm 1.96 \left(\text{var}(1 - E\hat{\rho}^2)^{\frac{1}{3}} \right)^{0.5} \right)^3 \right] \tag{3.44}$$

As an illustration of the methodology proposed by Schroeder and Hakstian (1990), and as an indication of the variability of the generalizability coefficient, we estimate the interval around the generalizability coefficient for one of the problems discussed in Finn and Kayandé (1997), using only the balanced portion of the data.

We summarize the discussion on estimation methods with a classification of the point and interval estimation methods in terms of the situations under which each method is most appropriate. Table 3.1 presents a simple 2 X 2 classification on the nature of the data (balanced or unbalanced) and type of estimate (point or interval), and the summary of methods recommended.

Insert Table 3.1 about here

3.3 Empirical Illustration

In the preceding theory section of the paper, we reviewed the different methods of point and interval estimation of variance components. We now explore the empirical implications of the differences in methods of estimation on the optimal design of measurement, the questions we raised at the beginning of this study. We use the data collected by Finn and Kayandé (1997) to illustrate the differences because of two reasons. First, Finn and Kayandé's (1997) study is the only complete application of a generalizability theory based approach to measurement in marketing. Second, it provides a comparison standard, specially in the context of the impact of estimation methods on the optimality of designs generated from their study.

In this section, we first provide a brief description of the data collected by Finn and Kayandé (1997). Then, for a specific design included in their study, we estimate generalizability coefficients using variance components estimated by different methods for both balanced and unbalanced data. Such a comparison will indicate the extent to which the point estimation methods and balance in the data impact the estimated "quality" of information, as reflected in the generalizability coefficient. We also present optimal designs using variance components estimated from different methods, for a pre-specified generalizability coefficient of 0.90. This comparison gives an indication of the direct impact on the design of the optimal decision study, and therefore the cost of the decision study.

Then, we estimate intervals on variance components using the Satterthwaite method, Ting et al. method, and maximum likelihood methods. The first two methods are restricted to balanced data, because of the strong recommendation not to use the ANOVA method to estimate variance components with unbalanced data (note that both these methods use the mean squares obtained from the ANOVA method of estimation). The interval estimates from maximum likelihood methods are provided for both balanced and unbalanced data. The interval estimates of variance components may not provide a

conclusive answer to the question of whether the optimal designs are impacted by the variance around the estimates of variance components. This is because the optimality of designs is dependent on the ratio of variance components, and therefore would depend on the variability in this estimated ratio. We use Schroeder and Hakstian's (1990) method to estimate an interval for this ratio for one specific design. However, for other designs, including the "optimal" designs, the extent of variability in variance components will have to be taken as an approximate indication of the variability of the estimate of the ratio. This then would provide some indication of whether or not the optimality of designs generated by Finn and Kayandé's (1997) method is influenced by the interval estimates.

3.3.1 Brief Description of Data Collected by Finn and Kayandé (1997)

Data were collected by mail from 125 respondents on their service quality perceptions of 3 retail chains from each of 3 retail sectors. The respondents were asked to rate the service quality of the retail chains on 3 items, drawn randomly from the set of items contained in SERVQUAL (Parasuraman, Zeithaml, and Berry 1988), for each of 3 randomly chosen SERVQUAL dimensions, called aspects in this paper. Thus, the factors in the design were retail sectors (3 levels), retail chains (3 levels nested within 3 sectors), respondents (125 levels, crossed with all other factors), aspects (3 levels), and items (3 levels nested within aspects).Variance components were estimated by MIVQUE(0) and were used to estimate generalizability coefficients for various designs, and thereafter to estimate the optimal designs for different decision problems. Because we also compare methods on the balance in the data, we used the balanced subset of these data to estimate models on balanced data. This balanced subset, from 65 respondents who responded to every item for every retail chain, contains 76% of the data points contained in the full unbalanced data provided by the 125 respondents.

All other details of the data collection methodology are provided in Finn and Kayandé (1997).

3.3.2 Impact of Point Estimation Method on Optimal Designs

3.3.2.1 Impact on generalizability coefficient:

The different methods of estimation were used to estimate variance components, and these estimates were used to estimate generalizability coefficients for a single problem of comparing the service quality of 5 retail chains within any one retail sector. In addition, the design suggested to be optimal by Finn and Kayandé (1997) for achieving a generalizability coefficient of 0.90 was chosen for estimating the generalizability coefficient under each method. The design - 31 respondents, 1 item for each of 4 aspects - was suggested to be optimal by Finn and Kayandé (1997) using MIVQUE(0) estimates of variance components. These estimates in their paper were for the full unbalanced data; in this paper, we estimate the generalizability coefficients using different methods for both balanced and unbalanced data.

The variance components estimated by ANOVA, MIVQUE(0), ML and REML for the unbalanced data are shown in Table 3.2. One difference between the methods is that the ML methods produce estimates that are constrained to be non-negative, while the ANOVA and MIVQUE(0) method produced three negative estimates (retail sectors, sector by aspects, and sector by respondent by aspect). Note that we have set these estimates to 0 in the Table.

Insert Table 3.2 about here

There are differences among the specific estimates from each method; however, such differences have a significant impact only in terms of ratios of variance components. One common comparison standard is the percentage of total variance accounted for by each source of variance. There are no significant differences in percentage of total variance for most sources of variance across the four methods, with the exception of three sources for which there are noticeable, although minor, differences. The variance

component associated with retail chains accounts for 15% of total variance with ANOVA estimates, 14% with MIVQUE(0) estimates, but only 11% with ML and REML estimates. Thus, any coefficient formed with chains as the object of measurement should be smaller with ML and REML estimates than with ANOVA and MIVQUE(0) estimates, if the relative error variance is the same across the different methods. However, the relative error variance is also larger (as a percent of total variance) for ML and REML estimates, implying that there is a double effect to further reduce the generalizability coefficient for comparing chains with ML and REML estimates. This difference is reflected in the generalizability coefficients being lower for ML and REML estimates (0.867 and 0.875 respectively) than for ANOVA and MIVQUE(0) estimates (0.907 and 0.902 respectively).

The other two sources for which there are differences in percent of total variance across the methods are respondents (ranges from 16% with ANOVA estimates to 19% with ML estimates) and the interactions of chains and respondents (21% with ANOVA and MIVQUE(0) estimates and 23% with ML and REML estimates). The increase in percent of total variance for respondents with ML estimates is compensated for by a decrease in percent of total variance for the interaction of chains and respondents, a component of relative error for comparing respondent perceptions. Thus, the differences would have little impact on a G-coefficient for the comparison of respondent perceptions.

There are differences in the generalizability coefficient for comparing chains across different estimation methods; however, the differences are not significant enough to worry about. Thus, although the method of estimation has some impact on the generalizability coefficient, the impact does not seem strong enough to warrant major attention *in these data*. The differences across methods do, however, alert the attention of researchers to the possible impact of estimation method on generalizability.

The results from the balanced subset of the data are produced in Table 3.3. The variance components for ANOVA and MIVQUE(0) are identical, and these estimates are

different from ML and REML estimates for some sources of variances. However, notice how similar the estimates are for most sources of variance. In the theory section of this paper, we mentioned that REML estimates are identical to ANOVA estimates in the case of balanced data. We find however, that there are some differences across REML and ANOVA estimates. The reason is the presence of negative estimates with ANOVA; thus, the caveat to this property is important. If the estimates are negative with ANOVA, the solutions with REML and ANOVA should not be expected to be similar specially since REML constrains the estimates to be non-negative. There are hardly any differences in the percent of total variance for any source of variance across the four methods; thus, most ratios of variance components should be similar. This is reflected in the similarity of the generalizability coefficients across the four methods. We used the same design -31 respondents, 1 item for each of 4 aspects - as for the unbalanced design to estimate the generalizability coefficients for retail chains. The generalizability coefficient ranges from 0.855 with ML estimates to 0.865 with ANOVA and MIVQUE(0) estimates. Thus, the impact of method of estimation is minimal in terms of the estimate of generalizability coefficient from point estimates of variance components, in the context of balanced data. The differences are minor, although conspicuous, with unbalanced data.

Insert Table 3.3 about here

3.3.2.2 Impact on Optimal Designs:

The impact of different estimation methods on generalizability coefficients should be reflected in an impact on the optimality of the design of a decision study. That there were hardly any differences across methods for balanced data should be reflected in hardly any differences in optimal designs across different methods. However, we should expect to find some differences in optimal designs for unbalanced data, because there were some differences in the generalizability coefficient across the methods. In this

section, we report on differences in optimal designs across different methods for balanced and unbalanced data.

In addition, we extend the investigation to two other problems identified by Finn and Kayandé (1997), which lead to different objects of measurement. These are the problems of identifying priorities for quality improvement (or comparing aspects of service quality for a retail chain, Problem 2 of their paper) in a retail chain and determining customers' perceptions of service quality (or comparing the perceptions of respondents nested within retail chains, an adaptation of Problem 4 of their paper). We extended the investigation because there might be a different impact on the generalizability coefficient, and therefore optimal designs, for different problems.

For all problems examined in this paper, we used the cost function and optimization framework developed by Finn and Kayandé (1997). Estimates of variance component serve as an input to the optimization; different estimates served as input, based on the method of estimation.

Optimal designs, including the associated costs, for all three problems under different methods of estimation and balanced versus unbalanced data are given in Table 3.4. We now describe the differences for each problem sequentially. The first problem is that of comparing the service quality of 5 retail chains within any one retail sector. The facets of generalization are respondents, items, and aspects. The optimal design using ANOVA method on unbalanced data is the cheapest ($438) for Problem 1, across both balanced and unbalanced data. However, note that the ANOVA estimation method is strongly *not recommended* with unbalanced data. Thus, this lowest cost design may not really be all that optimal.

The optimal designs with balanced data are consistently more expensive than those with unbalanced data, not so much reflecting the lower survey response rates when balanced data are desired, but generally reflecting the lower generalizability coefficient for any given design. The most significant finding to take away from these results is that

the cost of designs can differ by a large percentage amount, even though there are no major differences in the generalizability coefficient. This is made quite obvious in this problem, where the optimal design with ML estimates is about 50% more expensive ($700 versus $479) than the optimal design with MIVQUE(0) estimates, cited as *the* optimal design by Finn and Kayandé (1997). If one is permitted comparisons across balanced and unbalanced data, then the optimal design with ML estimates for balanced data is about 65% more expensive than the optimal design with MIVQUE(0) estimates for unbalanced data.

Insert Table 3.4 about here

The results for the second problem of identifying priorities for quality improvement in a retail chain are similar to those for the first problem. The object of measurement for this problem is aspects for a fixed retail chain, and the facets of generalization are respondents and items. For unbalanced data, the costs of optimal designs range from $1154 with ANOVA estimates to $1670 with ML estimates. For balanced data, the costs range from $962 for ANOVA and MIVQUE(0) estimates to $1346 for ML estimates. The ML estimates result in the most expensive optimal designs for both problems, and both balanced and unbalanced data. The pattern for the second problem is not the same as that for the first problem, indicating that there is no empirical reason to expect any systematic differences in optimal designs across methods of estimation. In any case, there is no theoretical reason to expect any systematic differences.

The third problem is an adaptation of Problem 4 from Finn and Kayandé's (1997) study. The comparison of the service quality perceptions of 30 respondents nested within 5 retail chains is considered in this problem. The object of measurement is respondents nested within retail chains; the facets of generalization are items and aspects. The optimal

designs for this problem are identical across all methods of estimation and balanced versus unbalanced data. The optimal design costs $1470 and involves asking customers to respond to 1 item for each of 9 aspects of service quality.

3.3.3 Impact of Interval Estimation on Optimal Designs

In this section, we report on the empirical interval estimates of variance components, and their possible impact on the optimal design of a decision study. We first report the results of methods that use mean squares from ANOVA, and therefore restrict the discussion to balanced data. Subsequently, we report maximum likelihood interval estimates for both balanced and unbalanced data. Finally, we discuss the interval estimate of a generalizability coefficient using Schroeder and Hakstian's (1990) method. All estimates are derived using methods which have been described in detail in the theory part of this paper.

The 90% interval estimates using Satterthwaite's (1941, 1946) and Ting et al.'s (1990) methods are presented in Table 3.5. The variance components were estimated using the ANOVA method. The intervals produced by Satterthwaite's method are quite wide for most sources of variance. For example, the lower and upper bounds for the estimate of the variance component for retail chains are 0.259 and 1.679 respectively. Ting et al. 's (1990) method produces intervals that are wider than those with Satterthwaite's method, consistently for each source of variance except the error variance for which the intervals are identical with both methods. Note that intervals estimated by Ting et al.'s (1990) method, although wider, are supposed to be more conservative relative to the more liberal intervals produced by Satterthwaite's method (Burdick and Graybill 1992).

Insert Table 3.5 about here

The confidence intervals using maximum likelihood methods (both ML and REML) are estimated using the asymptotic dispersion matrix of the estimates of variance components and the normal distribution assumption for all effects. The intervals are presented for both balanced and unbalanced data in Table 3.6. Most intervals are again wide, indicating that the method of estimating intervals matters little in terms of the conclusion that the intervals around variance component estimates are too wide to provide reasonable confidence in the optimal designs derived using the point estimates. A number of intervals include negative lower bounds, indicating that the variance component estimate is not significantly different from zero. For example, the lower bound of the interval estimate for aspects is consistently negative across all sets of data and methods of interval estimation. It simply implies that conclusions drawn about optimal designs for identifying priorities for quality improvement (aspects as the object of measurement) are suspect, because the variance component associated with aspects is not significantly different from zero. A similar conclusion can be drawn for items. The lower bound for the variance component associated with retail chains is just above zero indicating that an interval with higher confidence (say, 95%) would probably indicate that the variance component for retail chains is not significantly different from zero. Thus, all optimal designs for problem 1 in Finn and Kayandé's (1997) study are susceptible to be rendered sub-optimal because of the variability and/or non-significance of the variance component estimate for retail chains.

Insert Table 3.6 about here

3.3.3.1 Confidence Interval on the Estimate of Generalizability Coefficient

The method suggested by Schroeder and Hakstian (1990) was used to obtain an interval estimate on the generalizability coefficient for a decision study that uses the same design as a generalizability study. The mean squares from the ANOVA estimates were

used to derive the interval estimate for the generalizability coefficient to compare retail chains, with the design that used 65 respondents rating 3 retail chains from each of 3 retail sectors on 3 service quality items from each of 3 aspects. The point estimate of the generalizability coefficient for this design, using the ANOVA estimates from the balanced data, is 0.897. Using the method recommended by Schroeder and Hakstian (1990), the 90% interval on this estimate is estimated to be [0.73 0.97]. This wide interval implies that the variability in the mean squares, and therefore variance components, can be expected to have a large negative impact on the confidence that can be placed in a design remaining optimal for multiple decision studies.

3.4 Discussion and Conclusions

In this paper, we set out to answer two important questions. First, does the point estimation method used in a study have an impact on the optimality of designs using Finn and Kayandé's (1997) procedure? Second, we asked whether the interval estimates of variance components and generalizability coefficients are such that the confidence placed in optimal designs may be suspect. In the process of answering these questions, we raised several other questions, some of which we answer in the paper and some of which we leave for future research. To answer the first question, we had to review the different point estimation methods available to estimate variance components, understand the statistics of each method, and apply the same to the empirical data collected by Finn and Kayandé (1997). To answer the second question, we had to review the available methods to form intervals around estimates of variance components, in the process identifying most of the issues in forming such intervals, understand the statistics underlying each method, and apply the understanding to estimate intervals for the empirical data.

The contribution of this paper, thus, can be viewed as two fold. First, we attempted to answer the two important questions raised upfront. Second, the paper provides a concise summary of the statistics of extant point and interval estimation methods that were used to answer the questions, and also provides summary

recommendations and issues, such as in Table 3.1, related to the methods and their application. We now discuss the answers to the two important questions.

The impact of the point estimation method on optimal designs was clearly dependent on the balance in the data. The impact is quite minimal with balanced data; with no negative estimates, there is no impact at all. The impact is stronger with unbalanced data. Maximum likelihood methods seem to lead to more expensive designs; however, Searle, Casella, and McCulloch (1992) recommend more frequent use of maximum likelihood procedures because of the desirable properties, such as consistency and asymptotic normality, of the estimates, specially in the context of unbalanced data. In addition, the asymptotic sampling dispersion matrix of the estimators is also known, which allows for easy interval estimation compared with methods such as ANOVA and MINQUE. Thus, the more expensive designs with maximum likelihood methods seem to be compensated for by the desirable properties of the estimates used to derive the designs. ANOVA method should not be used with unbalanced data, and so the fact that ANOVA estimates lead to a lower cost design, at least in our study, is of little comfort when there are unbalanced data.

There are no major systematic differences in the impact of method across different objects of measurement. In essence, the results show us that the cost of designs with estimates from different methods depends in part on the object of measurement. For example, the optimal designs for respondents (nested within chain) as object of measurement were the same across all methods and balance in data. On the other hand, the designs for the other problems were different across different estimation methods. One conclusion that can be made is that the optimal designs for ANOVA and MIVQUE(0) estimates with balanced data are identical across all problems, because the estimates of variance components are identical for the two methods.

We have not compared computational speed of the different estimation methods in this paper. With these data, MIVQUE(0) is unquestionably the best on computational

speed, with balanced *and* unbalanced data. Maximum likelihood methods are most cumbersome in time and power, and therefore their use is inhibited specially for large problems such as the one explored in this paper. More work is required on this issue, because the speed of computation may dictate the "total cost" of each method.

The answer to the second important question sought to be answered by this study raises several issues. Such a wide range for most of the interval estimates unequivocally indicates that the confidence intervals around the estimates of variance components are too wide to provide sufficient confidence in the generalizabilty coefficients and optimal designs derived by Finn and Kayandé (1997). In their paper, they suggested that most of the estimated non-negative variance components were significantly different from zero; however, they did not provide confidence intervals for the estimates. The estimates of confidence intervals in our paper provide the argument to treat the results of Finn and Kayandé (1997) with extreme caution. This does not mean that Finn and Kayandé's (1997) methodological framework is flawed because its application produces wide interval estimates. We continue to strongly support their methodological framework, simply because it is correct compared to classical methods. However, they appear to have seriously underestimated the number of levels of facets that are needed in a generalizability study to be able to rely on the conclusions as to the optimal design of the decision studies. What we call for is more investigation into the manner in which the application of Finn and Kayandé's (1997) can be improved.

As is obvious from the wide interval obtained for the generalizability coefficient using Schroeder and Hakstian's (1990) method, the confidence interval around the generalizability coefficient is quite wide, even with such a specific study that has a larger number of levels for each facet than would be typically found in a decision study. It implies two alternative conclusions, the choice of which to adopt is an open issue. The first conclusion is that not much confidence should be placed in the point estimates of generalizability coefficient, and therefore the optimal designs generated by using point

estimates of generalizability coefficients be treated with caution. The alternative conclusion is that there must be ways to improve the interval estimates of the generalizability coefficients. This essentially implies that researchers need to embark on projects that will lead to reduced interval width, thus increasing the confidence that they may place in the optimal designs.

Footnotes

1. A confidence interval is considered to be liberal if the empirical probability that the parameter of interest lies between the upper and lower bound is lower than the desired probability (Burdick and Graybill 1992). Alternatively, an interval is conservative if the empirical probability that the parameter of interest lies between the upper and lower bound is greater than the desired probability.

Table 3.1
Comparison of Alternative Methods for Estimation of Variance Components

	Point Estimation	Interval Estimation
Balanced data	• ANOVA methods produce minimum variance unbiased estimates that can be negative. • ML and REML produce non-negative estimates with desirable properties. • ML estimates are biased.	• Ting et al. (1990) provide a general procedure to produce approximate intervals with confidence coefficients close to true confidence coefficients. • Satterthwaite's procedure not recommended when there are large differences in sample sizes across effects. • The dispersion matrix of ML or REML estimates can be used to construct intervals; however, large-sample property is a potential problem.
Unbalanced data	• ML and REML yield non-negative estimates with desirable properties except biasedness. • MINQUE methods yield estimates that may be negative. • ANOVA methods *not* recommended.	• Methods not well defined, although approximations can be found for some designs in Burdick and Graybill (1992). • The dispersion matrix of ML or REML estimates can be used to construct intervals; however, large-sample property is a potential problem.

Table 3.2: Unbalanced data, point estimates, and generalizability coefficients for a 31 respondents, 4 aspects, 1 item decision study to compare retail chains

	Methods of Estimation			
Source of Variation	**ANOVA**	**MIVQUE(0)**	**ML**	**REML**
retail chain	0.745[a]	0.694	0.500	0.534
retail sector	0[b]	0[b]	0.000	0.000
respondents	0.786	0.850	0.877	0.878
aspects	0.198	0.193	0.133	0.175
items	0.073	0.074	0.075	0.075
sector by respondents	0.066	0.016	0.004	0.004
sector by aspects	0[b]	0[b]	0.000	0.000
sector by items	0.017	0.018	0.018	0.018
chain by respondents[c]	1.040	1.053	1.070	1.070
chain by aspects	0.093	0.093	0.091	0.091
chain by items	0.030	0.033	0.033	0.033
respondents by aspects	0.278	0.260	0.249	0.249
respondents by item	0.166	0.158	0.162	0.162
sector by respondents by aspects	0[b]	0[b]	0.002	0.002
sector by respondents by items	0.080	0.107	0.098	0.098
chain by respondents by aspects	0.594	0.633	0.603	0.603
error	0.775	0.752	0.758	0.758
Relative error variance	0.077	0.075	0.077	0.076
G-coefficient for chains	0.907	0.902	0.867	0.875

[a] All numbers truncated to 3 decimal places. Thus, a variance component of 0.000 does not necessarily imply that it is zero.

[b] Negative estimate set to 0.

[c] The variance components for the interactions of the object of measurement (chains) with the facets of generalization are in italics.

Table 3.3: Balanced data, point estimates, and generalizability coefficients for a 31 respondents, 4 aspects, 1 item decision study to compare retail chains

	Methods of Estimation			
Source of Variation	**ANOVA**	**MIVQUE(0)**	**ML**	**REML**
retail chain	0.520[a]	0.520	0.451	0.478
retail sector	0[b]	0[b]	0.000	0.000
respondents	0.716	0.716	0.706	0.708
aspects	0.222	0.222	0.162	0.216
items	0.076	0.076	0.076	0.076
sector by respondents	0[b]	0[b]	0.000	0.000
sector by aspects	0[b]	0[b]	0.000	0.000
sector by items	0.011	0.011	0.011	0.011
chain by respondents[c]	1.184	1.184	1.163	1.163
chain by aspects	0.103	0.103	0.089	0.089
chain by items	0.029	0.029	0.029	0.029
respondents by aspects	0.287	0.287	0.288	0.287
respondents by item	0.164	0.164	0.164	0.164
sector by respondents by aspects	0.058	0.058	0.056	0.056
sector by respondents by items	0.034	0.034	0.034	0.034
chain by respondents by aspects	0.514	0.514	0.515	0.515
error	0.681	0.681	0.681	0.681
Relative error variance	0.081	0.081	0.077	0.077
G-coefficient for chains	0.865	0.865	0.855	0.862

[a] All numbers truncated to 3 decimal places. Thus, a variance component of 0.000 does not necessarily imply that it is zero.
[b] Negative estimate set to 0.
[c] The variance components for the interactions of the object of measurement (chains) with the facets of generalization are in italics.

Table 3.4: Impact of Estimation Method on Optimal Designs for Problems 1, 2, and 4 from Finn and Kayandé (1997)

Problem 1: Benchmarking Retail Chains within a Retail Sector (Scaling service quality of 5 retail chains)

	Unbalanced Data				Balanced Data			
	ANOVA	MIVQUE	ML	REML	ANOVA	MIVQUE	ML	REML
Number of Respondents	36	31	45	39	46	46	54	49
Number of Items	1	1	1	1	1	1	1	1
Number of Aspects	3	4	5	5	5	5	5	5
Cost ($)	438	479	700	640	710	710	790	740

Problem 2: Identifying Priorities for Quality Improvement (Scaling 5 aspects of 1 retail chain)

Number of Respondents	67	64	78	63	51	51	64	57
Number of Items	7	8	10	9	7	7	9	7
Cost ($)	1154	1232	1670	1332	962	962	1346	1034

Problem 4: Determining Customers' Perceptions of a Chain's Quality

(Scaling service quality perceptions of 30 respondents for each of 5 retail chains)

1 item, 9 aspects, and a cost of $1470 with every estimation method

Table 3.5: Two-sided 90% confidence intervals on ANOVA estimates of variance components from balanced data

Source	$\hat{\sigma}^2$	Satterthwaite's Method		Ting et al.'s Method	
		Lower Bound	Upper Bound	Lower Bound	Upper Bound
retail sector	-0.055	na[a]	na	-0.570	2.365
retail chain	0.520	0.259	1.679	0.214	2.015
respondents	0.716	0.583	0.904	0.477	1.180
aspects	0.222	0.074	4.322	0.041	5.032
items	0.076	0.038	0.246	0.030	0.276
sector by respondents	-0.031	na	na	-0.137	0.154
sector by aspects	-0.022	na	na	-0.078	0.098
sector by items	0.011	0.008	0.018	-0.002	0.046
chain by respondents	1.184	1.072	1.315	1.015	1.388
chain by aspects	0.103	0.064	0.202	0.049	0.258
chain by items	0.029	0.027	0.030	0.016	0.035
respondents by aspects	0.287	0.251	0.332	0.205	0.470
respondents by item	0.164	0.154	0.176	0.132	0.203
sector by respondents by aspects	0.058	0.054	0.063	0.010	0.169
sector by respondents by items	0.034	0.032	0.035	0.003	0.067
chain by respondents by aspects	0.514	0.493	0.536	0.446	0.591
error	0.681	0.649	0.715	0.649	0.715

[a] *na* indicates non-applicable, because confidence intervals are not defined on negative estimates with Satterthwaite's method..

Table 3.6: Two-Sided 90% Confidence Intervals on Maximum Likelihood Estimates of Variance Components from Balanced and Unbalanced data

	Balanced data						Unbalanced data					
	ML			REML			ML			REML		
Source	$\hat{\sigma}^2$	Lower Bound	Upper Bound	$\hat{\sigma}^2$	Lower Bound	Upper Bound	$\hat{\sigma}^2$	Lower Bound	Upper Bound	$\hat{\sigma}^2$	Lower Bound	Upper Bound
retail sector	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
retail chain	0.451	0.050	0.852	0.478	0.039	0.918	0.500	0.098	0.901	0.534	0.053	1.015
respondents	0.706	0.418	0.994	0.708	0.419	0.997	0.877	0.589	1.165	0.878	0.616	1.139
aspects	0.162	-0.128	0.452	0.216	-0.213	0.645	0.133	-0.157	0.423	0.175	-0.183	0.532
items	0.076	-0.006	0.159	0.076	-0.006	0.159	0.075	-0.008	0.158	0.075	-0.008	0.158
sector by respondents	0.000	0.000	0.000	0.000	0.000	0.000	0.004	0.004	0.004	0.004	-0.112	0.121
sector by aspects	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
sector by items	0.011	-0.006	0.027	0.011	-0.006	0.027	0.018	0.001	0.035	0.018	-0.004	0.040
chain by respondents	1.163	1.015	1.310	1.163	1.015	1.310	1.070	0.923	1.218	1.070	0.919	1.222
chain by aspects	0.089	0.024	0.155	0.089	0.024	0.155	0.091	0.025	0.157	0.091	0.024	0.158
chain by items	0.029	0.014	0.044	0.029	0.014	0.044	0.033	0.018	0.048	0.033	0.017	0.050
respondents by aspects	0.288	0.194	0.381	0.287	0.194	0.381	0.249	0.155	0.342	0.249	0.179	0.318
respondents by item	0.164	0.134	0.195	0.164	0.134	0.195	0.162	0.131	0.193	0.162	0.132	0.192
sector by respondents by aspects	0.056	0.006	0.106	0.056	0.006	0.106	0.002	-0.049	0.052	0.002	-0.051	0.054
sector by respondents by items	0.034	0.009	0.058	0.034	0.009	0.058	0.098	0.074	0.123	0.098	0.066	0.131
chain by respondents by aspects	0.515	0.452	0.578	0.515	0.452	0.578	0.603	0.540	0.666	0.603	0.536	0.671
error	0.681	0.648	0.714	0.681	0.648	0.714	0.758	0.725	0.791	0.758	0.723	0.793

Bibliography

Blischke, W.R. (1968), "Variances of Moment Estimators of Variance Components in the Unbalanced R-way Classification," Biometrics, 24, 527-540.

Brennan, Robert (1983), Elements of Generalizability Theory. Iowa City, Iowa: ACT Publications.

Burdick, Richard K and Franklin A. Graybill (1992), Confidence Intervals on Variance Components. New York: Marcel Dekker.

Cronbach, Lee J., Goldine C. Gleser, Harinder Nanda, and Nageswari Rajaratnam (1972) The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles. New York: John Wiley & Sons.

Finn, Adam and Ujwal Kayandé (1997), "Reliability Assessment and Optimization of Marketing Measurement," Journal of Marketing Research, 34 (2), 262-275.

Graybill, F.A. (1954), "On Quadratic Estimates of Variance Components," Annals of Mathematical Statistics, 25, 367-372.

_____ and C.M. Wang (1980), "Confidence Intervals on Nonnegative Linear Combinations of Variances," Journal of American Statistical Association, 75, 8869-873.

_____ and R.A. Hultquist (1961), "Theorems Concerning Eisenhart's Model II," Annals of Mathematical Statistics, 32, 261-269.

_____ and A.W. Wortham (1956), "A Note on Uniformly Best Unbiased Estimators of Variance Components," Journal of American Statistical Association, 51, 261-268.

Hartley, H.O. and J. N. K. Rao (1967), "Maximum Likelihood Estimation for Mixed Analysis of Variance," Biometrika, 54, 93-108.

_____, J. N. K. Rao, and L. Lamotte (1978), "A Simple Synthesis-Based Method of Variance Component Estimation," Biometrics, 34, 233-244.

Khuri, A.I. and Hardeo Sahai (1985), "Variance Components Analysis: A Selective Literature Survey," International Statistical Review, 53, 3, 279-300.

LaMotte, L.R. (1970), "A Class of Estimators of Variance Components," Technical Report 10, Department of Statistics, University of Kentucky, Lexington, Kentucky.

_____ (1971), "Locally Best Quadratic Estimators of Variance Components," Technical Report 22, Department of Statistics, University of Kentucky, Lexington, Kentucky.

_____ (1972), "Notes on the Covariance Matrix of a Random, Nested ANOVA Model," Annals of Mathematical Statistics, 43, 659-662..

Parasuraman, A., Valarie A. Zeithaml, and Leonard L. Berry (1988), "SERVQUAL: A Multiple-Item Scale for Measuring Consumer Perceptions of Service Quality," Journal of Retailing, 64 (Spring), 12-40.

Patterson, H.D. and R. Thompson (1971), "Recovery of Inter-Block Information when Block Sizes are Unequal," Biometrika, 58, 545-554.

Paulson, E. (1942), "An Approximate Normalization of the Analysis of Variance Distribution," Annals of Mathematical Statistics, 13, 233-235.

Rao, C.R. (1970), "Estimation of Heteroscedastic Variance in Linear Models," Journal of American Statistical Association, 65, 161-172.

_____ (1971a), "Estimation of Variance and Covariance Components - MINQUE Theory," Journal of Multivariate Analysis, 1, 257-275.

_____ (1971b), "Minimum Variance Quadratic Unbiased Estimation of Variance Components," Journal of Multivariate Analysis, 1, 445-456..

_____ (1972), "Estimation of Variance and Covariance Components in Linear Models," Journal of American Statistical Association, 67, 112-115.

_____ (1973), Linear Statistical Inference and its Applications, 2nd ed., John Wiley and Sons: New York.

Satterthwaite, F.E. (1941), "Synthesis of Variance," Psychometrika, 6, 309-316.

_____ (1946), "An Approximate Distribution of Estimates of Variance Components," Biometrics Bulletin, Vol. 2, 110-114.

Schroeder, Marsha L. and A. Ralph Hakstian (1990), "Inferential Procedures for Multifaceted Coefficients of Generalizability," Psychometrika, 55, 3, 429-447.

Searle, Shayle R. (1987), Linear Models for Unbalanced Data. New York: John Wiley and Sons.

_____, George Casella, and Charles E. McCulloch (1992), Variance Components. New York: John Wiley and Sons.

Shavelson, Richard J. and Noreen M. Webb (1981), "Generalizability Theory: 1973-1980," British Journal of Mathematical and Statistical Psychology, 34 (November), 133-166.

Ting, Naitee, Richard K. Burdick, Franklin A. Graybill, S. Jeyaratnam, and Tai-Fang C. Lu (1990), "Confidence Intervals on Linear Combinations of Variance Components that are Unrestricted in Sign," Journal of Statistical Computation and Simulation, 35, 135-143.

Welch, B.L. (1956), "On Linear Combinations of Several Variances," Journal of American Statistical Association, 51, 132-148.

Wonnacott, T. (1987), "Confidence Intervals or Hypothesis Tests?," Journal of Applied Statistics, 14, 195-201.

Chapter 4

Design and Reliability Assessment of Measurement of Multidimensional Marketing Constructs

The assessment of reliability of information gathered from marketing surveys is a necessary activity because of the potential impact of measurement error on the conclusions derived from the information. In the marketing literature, researchers have commonly used Cronbach's alpha (Peter 1979) as the measure of reliability. Rentz (1987, 1988) presented reasons, based on generalizability theory (Cronbach et al. 1972), for discontinuing the use of Cronbach's alpha as a measure of reliability. More recently, Finn and Kayandé (1997a) demonstrated the usefulness of the generalizability theory approach for assessing generalizability (a general form of reliability) and optimizing future studies by taking into account the purpose of measurement.

In Finn and Kayandé's (1997a) paper, the discussion of the measurement of marketing objects was implicitly restricted to single unidimensional constructs. Any discussion of multiple dimensions of a construct was restricted to an assumption that the dimensions are exchangeable with each other. However, neither do constructs necessarily have to be unidimensional nor are the dimensions necessarily exchangeable with each other; indeed many of the important characteristics and constructs measured in marketing are multidimensional. Finn and Kayandé (1997b) find that about 65% of all scales developed to measure marketing constructs are multidimensional. In light of this finding, two important questions arise. The first question that needs to be answered is whether the reliability assessment methods are different for multidimensional constructs than those for unidimensional constructs. Second, in the context of Finn and Kayandé's (1997a) methodology for designing optimal future measurement, does the multidimensionality of measured constructs have an impact on optimal design of future measurement?

In this paper, we argue that the answers to these questions are very much dependent on the way in which the multidimensional information is used by managers

and researchers. The most important distinction is in terms of deciding whether to analyze data separately for each dimension, thus treating the dimensions as distinct unidimensional constructs, or combine the dimensions into *a composite or an index*, with weights for each dimension determined by some specific criterion. The distinction is explained best with an example. Consider the problem of assessing the performance of managers of restaurants in a multi-restaurant chain such McDonald's or Wendy's, for determining a performance-based incentive. Performance may be determined on several dimensions such as the quality of service provided by the restaurant and quality of food (products) served by the restaurant, as evaluated by customers or mystery shoppers (Finn and Kayandé 1997c). The reliability of the measurement of performance on these dimensions can be assessed separately for each dimension. The most efficient designs for measuring each dimension separately can be generated by the methods suggested by Finn and Kayandé (1997a). A separate assessment and optimization on each dimension can help provide reliable information on each dimension, which can then be used in a diagnostic sense to improve performance on each dimension. However, most firms provide incentives based on some combination of performance scores on these multiple dimensions. Clearly, both dimensions play an important role in determining the attractiveness of a restaurant for its customers, and therefore sales and market share. An easy method to combine the scores from each dimension is to create a linear combination with equal weights, thereby suggesting that both dimensions are important, but can be traded-off with each other. However, would the dimensions be given equal weight even though it is known that the objects of measurement, in this case restaurants, do not differ much on one of the dimensions, say product quality? Or more importantly, would the weights be equal even if it is known that the product quality and service quality have been measured with differential reliabilities? Finally, would the ideal weights still be equal if, at the same reliability level, the cost of measurement with unequal weights is lower than that with equal weights?

The preceding discussion is indicative of the different ways in which a firm may wish to use the information on a multidimensional construct such as quality or performance. It is also indicative of the fact that a composite score, a combination of scores on multiple dimensions, results in a gain in parsimony at the potential cost of reliable diagnostic information. The gain in parsimony may also result in a gain in terms of lower costs of data collection. The discussion also suggests that a composite score can be constructed in different ways, depending on the choice of weights. The weights in a composite can be pre-determined (equal weights or gain/difference scores), or the weights may be determined using some criterion such as the variation of the object of measurement on each dimension, reliability of the measurement of each dimension, and/or cost of measurement on each dimension.

In this paper, we focus on the methods to determine the reliability of composite scores and optimization of measurement in situation which require composite scores. We will also provide a comparison to the case when composites are not formed and dimensions are treated as distinct constructs, thus illustrating the costs of obtaining reliable diagnostic information. However, the focus will be on cases when a firm needs to combine the information from different dimensions. Some examples of such cases will be provided, while the case of retail quality will be used as an empirical illustration.

In the next section, we review the method currently used for assessing reliability of the composite scores formed from multidimensional measures. Then, we present examples of cases where there might arise a need for forming composite scores on multiple dimensions. Then we present a methodology to assess the generalizability of composite scores, using equal *or* unequal weights. We also suggest a method to determine the weights that lead to maximal generalizabilty for the composite score. Next, we incorporate this method into the optimization framework developed by Finn and Kayandé (1997a) for unidimensional constructs. We then provide an empirical illustration of the methodological framework, using retail store quality as the illustrative context.

Finally, we discuss the implications of the framework, summarize the contribution of the paper, and suggest avenues for future research.

4.1 Current Methods to Assess Reliability of Multidimensional Constructs

Classical reliability theory methods typically recommend the evaluation of the reliability of multidimensional measurement by first assessing the classical reliability coefficient (typically Cronbach's alpha) for each dimension separately, and then using Nunnally's (1978) formula for the reliability of a linear combination (for examples of this procedure, see Parasuraman, Berry, and Zeithaml 1988 and Bienstock, Mentzer, Bird 1997). The formula for the reliability of a difference score (Peter, Churchill, and Brown 1993) is derived using Nunnally's formula. Nunnally's method has two important limitations. First, it follows classical reliability theory and therefore all limitations of that theory, detailed in Finn and Kayandé (1997a), hold for multidimensional measurement. The most important of these limitations is the continued focus on scaling of respondents, most often not the purpose of measurement. Second, Nunnally's (1978) formula ignores the possibility of covariances across dimensions. Although a multivariate analogue of classical reliability was proposed earlier by Bock (1963, 1966) and Conger and Lipshitz (1974), their methods have limited utility because of their continued focus on single-faceted measurement.

Another literature on reliability comes from the factor analysis tradition. Cronbach's alpha requires the items to be strictly parallel or tau-equivalent. Such a strong requirement encouraged the development of Coefficient theta (Armor 1974) and Coefficient omega (Heise and Bohrnstedt 1970), which are estimated using the eigenvalues and communalities from a factor analysis model. Coefficient theta is derived from the principal components model, whereas coefficient omega is derived from the common factor analysis model. Both models, and therefore measures of reliability, focus on scaling of respondents and single-faceted measurement. Therefore, from our multi-faceted perspective, they also suffer from similar limitations as Cronbach's alpha.

4.2 Composite Scores in Marketing

Multidimensional information can be used in different ways to suit the managerial purposes of measurement. The multiple dimensions could be treated as separate dimensions and the reliability of scaling the object of measurement could be assessed separately for each dimension. This usage assumes that there is no need to quantitatively combine the information provided by each dimension into a composite. Some managerial purposes might require the information from each dimension to be combined in some way. This combination can be achieved by appropriately weighting each dimension. The weights may be equal, pre-determined (in case of a difference score), or may depend on some criterion such as the estimated reliability for each dimension. We now present three illustrative examples of cases in which a need to form composite scores may arise.

Consider the measurement of the quality of a retail outlet. This is a general version of example of the restaurant manager's performance evaluation. The quality of a retail outlet can be represented by the quality of both *the service provided* by and *the products sold* at the outlet. Thus, service quality and product quality constitute two distinct dimensions of retail store quality. A consumer's evaluation of the quality of a retail store will be dependent on the performance of the retailer on both dimensions. These dimensions of service quality and product quality are not necessarily highly correlated across all retail outlets. A retailer that carries only high quality products might simultaneously provide poor service quality, and vice versa. Therefore, the consumer's choice of a retail outlet might also depend on a trade-off between these two "attributes" or dimensions of a retail outlet.

It might be of managerial interest to measure the quality on both dimensions to get insights into how their outlets are performing on each dimension. In such cases, a comparison of retail outlets on their quality must focus on each dimension separately. A manager of a retail chain might also be interested in distinguishing *between overall best and worst* retail outlets; this requires a simultaneous consideration of both dimensions.

Additionally, the manager might be interested in finding out which dimension most impacts this discrimination between retail outlets. In a study of retail service quality, Dabholkar et al. (1996) argue that practitioners are often interested in overall service quality as well as dimensions of service quality. They suggest that using multiple items to develop a measure of overall service quality might be better than a single-item overall measure. Thus, the question becomes, "How can the information from multiple dimensions be best combined?" The quality of the retail outlet could be represented by a simple average of the scores obtained for the two dimensions of qualities. This would be a simple, equally weighted average of the scores on each dimension. However, the firm might also be interested in creating a weighted average that *discriminates best* between the retail outlets. In such a situation, it is essential to develop a method to estimate the weights that will lead to optimal discrimination between retail outlets.

The evaluation of the performance of salespeople in a firm is a second example of multidimensional measurement. It is well-known that objective measures of sales performance, such as dollar or unit sales, account for only a small fraction (5% to 8%) of the variance in the managerial evaluations of salesperson performance (Weitz 1978, Behrman and Perreault 1982). MacKenzie, Podsakoff, and Fetter (1993) show that the "organizational citizenship behavior" exhibited by salespeople has an impact on the managerial evaluations of salesperson performance. The organizational citizenship behavior is a construct composed of multiple dimensions such as altruism, civic virtues, sportsmanship, and conscientiousness. The authors show that these dimensions impact the managerial evaluation of performance of salespeople. Although descriptive in nature, the study's findings can be used by a sales manager to devise a method to evaluate salespeople on these different dimensions. The measurement procedure can be designed optimally within each dimension for developing scales of adequate reliability. In addition, the manager might wish to construct a composite of the dimensions, to represent overall citizenship behavior. The composite can be equally weighted, or constructed in such a

manner as to provide the most generalizable scaling of salespeople. For a company with a large salesforce, the number of variables to be used to determine salesperson performance may have to be restricted and therefore, multiple dimensions may need to be collapsed into a single composite, without any loss of generalizability of information and with potential benefits of lower costs of measurement.

Consider a third example of a bank that is attempting to assess the creditworthiness of its customers who apply for bank loan. The creditworthiness of an individual is a multidimensional construct; however, the assessment made by the bank should be an overall composite score, in order to ensure similar standards across all customers. Such a composite can be formed by weighting the different dimensions according to the importance of each dimension to the bank. Alternatively, the weights could be determined in such a way that it maximizes the generalizability of the composite scores. Note that generalizability is an estimate of the intra-class correlation coefficient that measures the extent to which observed scores are correlated with true scores (or universe scores, as they are called in generalizability theory terminology).

A composite score, whether formed by equal or unequal weighting, has to be evaluated on its reliability, because an unreliable composite score will have the same negative impact as an unreliable individual dimension score. But can the reliability of a composite score be evaluated in a fashion similar to that of an individual dimension? In the next section, we discuss a method to assess the reliability of a composite score formed from a multidimensional construct.

4.3 Assessment of Generalizability of a Composite Score

The generalizability coefficient for each dimension can be separately assessed using methods presented in Finn and Kayandé (1997a). Thus, all methods presented in their paper can be applied to a multidimensional construct, assuming that there is no need to form a composite score from the individual dimensions. The possibility that the decision maker might be interested in a *composite* of multiple dimensions calls for an

assessment of the generalizability of the measurement of the composite. The preceding section presented some examples of situation where managers might be interested in composite scores. In a retailing context, a composite quality score for a retail outlet can allow a researcher to examine the choices of customers, taking into account the trade-off between service and product quality. This composite may provide a better indication of an overall evaluation of the quality of the grocery chain than either dimension independently, because of the trade-offs inherent in consumer choice.

4.3.1 Multivariate Random Effects Model

The univariate generalizability coefficient, given in Finn and Kayandé (1997a), is derived from a univariate random effects model. The univariate random effects model is extended here to the multivariate case for the multi-dimensional measurement context. We use the example of measurement of the quality of grocery chains, where the quality is assumed to be two dimensional. Information is required by management on both the service quality provided by grocery chains and the quality of products available at the different grocery chains. We present only the two-dimensional extension of the univariate model as an illustration, without any loss of generality. While this is the simplest case, it can be easily extended to any number of dimensions. The multivariate random effects model for the two-dimensional case can be represented as,

(4.1)
$$X_{1ijk} = \mu_1 + \alpha_{1i} + \beta_{1j} + \gamma_{1k} + \alpha\beta_{1ij} + \alpha\gamma_{1ik} + \beta\gamma_{1jk} + \varepsilon_{1ijk}$$
$$X_{2ijk} = \mu_2 + \alpha_{2i} + \beta_{2j} + \gamma_{2k} + \alpha\beta_{2ij} + \alpha\gamma_{2ik} + \beta\gamma_{2jk} + \varepsilon_{2ijk}$$

for i = 1, 2, 3, ... , G grocery chains, j = 1, 2, 3, . . . , R respondents, k = 1, 2, 3, ... , K items, and N = G X R X K. μ_1 and μ_2 represent the overall means of service quality and product quality. α_{1i} and α_{2i} are the effects of grocery chain i on the two variables respectively, β_{1j} and β_{2j} are the effects of respondent j on the two variables, and so on. The random effects model conditions imply that,

$\text{var}(\alpha_1) = \sigma^2_{\alpha_1} I_G$, $\text{var}(\alpha_2) = \sigma^2_{\alpha_2} I_G$, $\text{var}(\beta_1) = \sigma^2_{\beta_1} I_R$, . . . , $\text{var}(\varepsilon_1) = \sigma^2_{\varepsilon_1} I_N$, and

$$var(\varepsilon_2) = \sigma^2_{\varepsilon_2} I_N,$$

and all effects within a dimension are independent of each other.

The covariance components for the effects are then given by,

$$cov(\alpha_1, \alpha'_2) = \tau_\alpha I_G,\ cov(\beta_1, \beta'_2) = \tau_\beta I_R, \ldots, cov(\varepsilon_1, \varepsilon'_2) = \tau_\varepsilon I_N$$

This is equivalent to the assumption that $cov(\alpha_{1i}, \alpha_{2i}) = \tau_\alpha$ for all i, but $cov(\alpha_{1i}, \alpha_{2i'}) = 0$ for all $i \neq i'$, and so on for every effect.

The multivariate variance components model for this two-dimensional fully crossed three-factor design is then given by,

$$\Sigma_X = \Sigma_\alpha + \Sigma_\beta + \Sigma_\gamma + \Sigma_{\alpha\beta} + \Sigma_{\alpha\gamma} + \Sigma_{\beta\gamma} + \Sigma_\varepsilon \tag{4.2}$$

where Σ_X is a 2 X 2 (representing the 2 dimensions) variance-covariance matrix of observations, Σ_α is the 2 X 2 variance-covariance components matrix of the effects due to grocery chains, Σ_β is the 2 X 2 variance-covariance components matrix of the effects due to respondents, . . . , and Σ_ε is the 2 X 2 variance-covariance components matrix of the error term.

4.3.2 Multivariate Generalizability Coefficient

The univariate generalizability coefficient can now be extended to the multivariate case by assuming that the multiple dimensions are to be used to form a composite score for each retail outlet. The multivariate coefficient (Joe and Woodward 1973, Woodward and Joe 1976) with grocery chains as the object of measurement is given by,

$$\rho = \frac{\mathbf{w}'\Sigma_\alpha \mathbf{w}}{\mathbf{w}'\Sigma_\alpha \mathbf{w} + \mathbf{w}'\left[\Sigma_{\text{relative error}}\right]\mathbf{w}} = \frac{\mathbf{w}'\Sigma_\alpha \mathbf{w}}{\mathbf{w}'\Sigma_\alpha \mathbf{w} + \mathbf{w}'\left[\Sigma_{\alpha\beta} + \Sigma_{\alpha\gamma} + \Sigma_\varepsilon\right]\mathbf{w}} \tag{4.3}$$

where ω' is a vector of weights to be used in constructing the composite score, and the elements of the relative error variance-covariance components matrix have been appropriately divided by the respective numbers of items and respondents. In the example of grocery chains, this coefficient evaluates the generalizability of composite scores, with appropriately chosen weights for service and product quality, for grocery chains.

4.3.4 Choice of weights

The weights can be chosen on the basis of theory or, in the absence of theory, can be equalized. Cronbach et al. (1972) suggest that composites scores can be formed for the purposes of examining the generalizability of difference or gain scores, commonly used in the assessment of student ability in educational psychology. Difference scores are simply composites with 1 and -1 as the weights for each dimension. Similarly, difference scores have been used prominently in work on service quality measurement by Parasuraman and his colleagues (Parasuraman, Zeithaml, and Berry 1988; Parasuraman, Berry and Zeithaml 1991). They use the difference between a respondent's perceptions and expectations to conceptualize service quality. Difference scores are commonly used in other areas of marketing, as evidenced in the review by Peter, Churchill, and Brown (1993). In all such cases, the weights are determined *a priori*.

An alternative view of Equation 4.3 suggests that it can be interpreted as an expression in which the weights can be selected to maximize the generalizability coefficient for any given design of interest for a subsequent decision study, as opposed to being determined purely by theory. The variance-covariance components for both the object of measurement and relative error are known, and therefore the weights can be chosen to maximize the generalizability coefficient (Joe and Woodward 1973). Choosing the weights to maximize the generalizability coefficient is equivalent to solving the following eigenvalue problem:

$$\left[\left[\Sigma_{\text{relative error}} + \Sigma_{\alpha}\right]^{-1} \Sigma_{\alpha} - \rho_s\right] \mathbf{w}_s = 0 \tag{4.4}$$

In this eigenvalue problem representation, the vector of weights ω_s is an eigenvector corresponding to an eigenvalue ρ_s of the matrix $\left[\Sigma_{\text{relative error}} + \Sigma_{\alpha}\right]^{-1} \Sigma_{\alpha}$. Generally, an additional constraint used in solving the eigenvalue problem is that $\omega'\omega = 1$, thus normalizing the eigenvector ω to have length 1.

Equation 4.4 shows that the choice of the eigenvector corresponding to the largest eigenvalue as the vector of weights ω will result in a composite of maximal

generalizability for any given design. Thus, instead of choosing the number of levels of sources of error in order to increase generalizability, the weights are chosen to maximize the generalizability of a composite score from any given design.

This representation also indicates the dimensions on which the object of measurement is best measured with a research design. The dimension with highest weight is the dimension on which the object is best discriminated. In other words, the signal to noise ratio is highest on that dimension. For example, in the two dimensional problem of comparing the quality of grocery chains, an optimal weight of 0.95 for service quality (which implies a weight of 0.31 for product quality, assuming a normalized eigenvector) implies that the *measurement procedure* is better at discriminating between grocery chains on service quality rather than the product quality of the chains. Note how this method recognizes that any measurement procedure is prone to error, and dimensions are weighted according to how well the dimension discriminates between the objects of measurement *as well as the quality of measurement of the variable* (alternatively, the extent to which it is measured with error).

4.3.5 Optimization of Multi-Dimensional Measurement Designs

The multivariate generalizability coefficient can be used in the procedure suggested in Finn and Kayandé (1997a) to optimize the design of future decision studies. Thus, we now include the cost of a measurement study as a factor to optimize the decision study. Following Finn and Kayandé (1997a), cost C of a decision study can be represented as,

(4.5) $$C = f(c_0, \tilde{C}, \tilde{N}, n_{object})$$

where c_0 is the fixed cost of the survey instrument, $\tilde{C}$ is a vector with elements representing the cost of an observation on each facet, $\tilde{N}$ is a vector with elements n_1, n_2, ...,n_F representing the number of levels of F facets of generalization, and n_{object} is the number of levels of the object of measurement.

The purpose of the optimization is to minimize the cost of the design subject to achieving a desired value g of the multivariate G-coefficient ρ given in Equation 4.4. Given a choice of weights **w**, the denominator in Equation 4.4 is a decreasing function in all F elements of $\tilde{N}$. Thus, the optimization problem can be formally stated as follows:

(4.6) $$\text{Minimize} \quad C = f(c_0, \tilde{C}, \tilde{N}, n_{object})$$
$$\tilde{N} = (n_1\ n_2\ .....\ n_F)$$

subject to,

1. $\rho \geq g$, the desired G-coefficient.
2. Each element of $\tilde{N}$ 1.
3. Each element of $\tilde{N}$ is an integer.

In addition to the choice of number of levels of the F facets of generalization, the choice of weights can also be included in this formulation. Thus, the minimization of cost is achieved by simultaneously choosing an optimal vector of weights for a composite score and an optimal design for a decision study. This simultaneous optimization should lead to a lower cost than a mere optimization using one of the two criteria.

4.3.6 Estimation of Variance-Covariance Components

Variance components can be estimated by several methods such as ANOVA, MIVQUE(0), and maximum likelihood (Searle, Casella, McCulloch 1992). Covariance components can be estimated by the same methods proposed for the estimation of variance components. The problem, however, is that standard statistical packages do not allow for estimation of covariance components. Also, the theory on estimating covariance components for unbalanced data is relatively sparse. Because the theory is sparse on this topic, we limit our focus to balanced data and designs. In balanced designs, the effects with covariance will be those that are common across the two dimensions. For example, both product quality and service quality of a single grocery chain is evaluated when forming a composite score. Thus, there is a covariance component between the effects of grocery chains on service quality and product quality. Similarly, the same respondents

might have evaluated the grocery chains on both dimensions. In such cases, we find a symmetric variance-covariance components matrix. However, it is unlikely that same items will be used to measure both dimensions. Typically, these dimensions are distinct and, therefore, they are measured with different items. The expected covariance for items across dimensions is then equal to zero. Therefore, the variance-covariance components matrix for such facets as items has variance components in the diagonal and all off diagonal elements are zero.

Searle, Casella, McCulloch (1992) provide an easy estimation method for covariance components. They show that for simple balanced designs (balanced and small number of factors) such as that of the grocery chain measurement, every covariance component can be represented by,

(4.7) $$\tau_{\alpha} = \text{cov}(\alpha_{1i}, \alpha_{2i}) = \frac{1}{2}\left(\sigma^2_{\alpha_{1+2}} - \sigma^2_{\alpha_1} - \sigma^2_{\alpha_2}\right)$$

where $\sigma^2_{\alpha_{1+2}}$ represents the components of variance for the variable $(X_1 + X_2)$, which is really a composite score. Further, they use general results from Searle and Rounsaville (1974) to show that,

(4.8) $$\hat{\tau}_{\alpha} = \frac{1}{2}\left(\hat{\sigma}^2_{\alpha_{1+2}} - \hat{\sigma}^2_{\alpha_1} - \hat{\sigma}^2_{\alpha_2}\right)$$

We use this general methodology for all covariance components that are estimated in the empirical study. More complex measurement problems, specially with unbalanced data, might not be adequately served by using this approximation. In such cases, it might be necessary to use maximum likelihood methods, which are beyond the scope of this paper.

The estimated variance-covariance components matrices can be negative definite because the methodology used to estimate them does not specifically preclude this possibility. There are two options to resolve this empirical possibility, which theoretically should not occur. First, Woodward and Joe (1976) and Webb and Shavelson (1981)

suggest that the researcher could simply set all the elements of the negative definite matrix to zero, thus eliminating the matrix from all calculations. However, this results in the elimination of information that can and should be used. In particularly troubling cases, it can mean few results, because of the possibility of having the variance-covariance components matrix associated with the object of measurement set to zero. The second possibility is to reconstruct the matrix with only the positive eigenvalues and associated eigenvectors. This can be viewed as the multivariate equivalent of setting a negative estimate of a variance component to zero (Terry Elrod 1996, personal communication). Thus, this approach leads to less loss of information than summarily setting the whole matrix to zero, and is to be generally preferred.

4.4 Empirical Illustration

The framework is illustrated with an application to the measurement of the quality of retail chains. Retail quality was chosen as the illustrative context because of the importance given by the marketing literature in recent years to the measurement of the quality of service and products. However, with the exception of Dabholkar et al. (1996), the quality of products and the quality of services have been treated independently, with any single study only concerning itself with one or the other. However, it should be obvious that the quality of a retail outlet is determined by both quality of products and quality of service provided. Thus, it makes sense to simultaneously consider both dimensions of retail quality. Our work differs from Dabholkar et al.'s study in terms of the primary focus and the object of measurement. Our primary focus is to illustrate a methodology whereas their focus was on assessing a model of retail service quality, which included the physical product aspect also. Second, the object of measurement in their study was always the respondent. Following Finn and Kayandé's (1997a) arguments, we suggest that the usefulness of a model that treats only the respondents as object of measurement is limited at best and misleading at worst.

Our study design included three randomly chosen dimension of SERVQUAL (Parasuraman, Berry, and Zeithaml 1991), which were combined to form a single dimension of service quality, and one additional dimension of product quality. The variance-covariance components for service quality dimensions were averaged to represent the single dimensions of quality of service provided by the retail chain. The items used in the questionnaire are shown in Appendix 4.1. We illustrate the procedure assuming that the quality of a retail outlet is represented by the two dimensions of service quality and product quality.

The four sources of variability, or facets, included in the multivariate G-study were, retail sectors (or type of retailer), retail chains, the items used to measure service quality, and consumers who responded to the questionnaire. The G-study had consumers evaluate the service quality and product quality provided by a total of nine retailers, three chosen at random from amongst the well known chains in three retail sectors. The specific chains were Eaton's, Wal-Mart and Zeller's from the department store sector, Dairy Queen, Kentucky Fried Chicken and McDonald's from the fast food sector, and Safeway, Save-on-Food, and Superstore from the grocery store sector. Each chain was evaluated on twelve items, nine items for the dimension of service quality and three items for product quality. The nine items for measuring service quality were drawn randomly from the SERVQUAL items representing three randomly chosen sub-dimensions of service quality, namely tangibles, responsiveness, and empathy. We chose to collapse the items around these three sub-dimensions into a single dimension for two reasons. First, Finn and Kayandé (1997a) showed that the variance component associated with these dimensions was negligible and therefore it is safe to assume that the dimensions are not distinct, when the purpose is to scale objects of measurement such as retail chains. Second, we chose two dimensions because of the resultant simplicity of exposition.

All other details of the data collection methodology, such as response rate and sample selection method, for this study are given in Finn and Kayandé's (1997a) paper.

In addition, because of the difficulties encountered with estimation on unbalanced data and because this study is illustrative only, we chose only those respondents who provided complete data. The balanced data from 65 respondents accounted for 76% of the full unbalanced data from 125 respondents.

4.5 Results

The variance-covariance components matrices for each random effect are provided in Table 4.1. There are no covariance components for items and any effect involving items because the items were not common across dimensions. On the other hand, effects involving respondents (but not items) have covariance terms because the same respondents responded to both product and service quality dimensions. An additional note is about the negative-definite matrices for retail sectors and items. Both matrices are negative-definite because of a negative variance component. Both matrices play no part in the estimation of generalizability coefficients, and hence the negative components have been truncated to zero.

Insert Table 4.1 about here

4.5.1 Univariate Results

Following Finn and Kayandé's (1997a) recommendations that the generalizability coefficient for each object of measurement should be estimated separately, we estimated the generalizability coefficients for the comparison of retail chains and respondents respectively. The generalizability coefficients for comparing retail chains and respondents are given at the bottom of Table 4.1 (these coefficients are estimated for single levels of each facet of generalization).

The generalizability coefficients for the comparison of both retail chains and respondents on service quality are higher than that for product quality. For the comparison of retail chains, the result was due to the variance component associated with

the interaction of items and chains being very small for service quality relative to the same component for product quality. In substantive terms, the finding implies that the scaling of chains did not depend on the different items measuring service quality whereas the scaling of chains depended very much on the different items measuring product quality. In other words, the service quality items had internal consistency (associated with the chains as object of measurement), whereas product quality items did not have internal consistency. For the comparison of respondents, although the variation in product quality perceptions among respondents was greater than the variation in service quality perceptions, the variation in service quality on each component of relative error variation was smaller than the same variation in product quality. Therefore, the generalizability coefficient for comparing respondents on service quality perceptions was greater than that for comparing respondents on product quality perceptions.

4.5.2 Optimal Designs for Decision Studies

We use the optimization framework provided in Equation 4.6 to find optimal designs to measure service quality and product quality for two different problems of comparing retail chains and comparing respondents. The cost function used in the optimization is given in Appendix 4.2. Optimal designs are derived for the univariate and multivariate case.

4.5.2.1 Problem 1: Comparison of Retail Chains

We consider the problem of comparing five retail chains in any one retail sector on the dimensions of service quality and product quality, and on a composite of the two dimensions. The weights for the composite are either equal or determined using Equations 4.4 and 4.6.

Table 4.2 provides the optimal designs in each case. The first design is for the measurement of service quality only; to achieve a generalizability coefficient of 0.90, 38 respondents will be required to provide ratings on 2 items measuring service quality. The cost of this decision study is $ 366. The second design illustrates the differences in

design, depending on the dimension being measured. To achieve a generalizability coefficient of 0.90 for measuring product quality, 79 respondents will have to respond to 9 items measuring product quality. The cost of this design is $ 1556, almost 4 times as much as that for measuring service quality.

Insert Table 4.2 about here

It is obvious that it is less efficient to conduct two separate studies for measuring service quality and product quality, when the same retail chains are being compared. The same respondents can be used to measure both service quality and product quality. The third design reflects this combination, where 79 respondents provide ratings on 1 item of service quality and 9 items of product quality for a total cost of $ 1685. The cost is cheaper than the combined cost of conducting two studies. Note how the number of service quality items reduced from 2 to 1, primarily because the increase in number of respondents was compensated for by a decrease in number of items.

These three designs cover the cases when the purpose is to separately analyze the information on each dimension for diagnostic purposes; thus, the designs provide independently generalizable measures for each dimension. We now consider designs which are optimal for composite scores. The first composite score considered here is the equally weighted composite of the two dimensions. A generalizability coefficient of 0.90 is expected to be achieved when 62 respondents provide ratings on 1 item measuring service quality and 3 items measuring product quality. The cost of such a decision study is $758, about half of the cost . a study which is expected to obtain independently generalizable scores for each dimension. The reduction in cost is substantial; however, this reduction is accompanied by a reduced diagnostic ability. While this design helps achieve a generalizable composite, it provides a measure for product quality that has a generalizability coefficient of 0.805, not sufficient for applied decision-making (Nunnally

1978). Thus, the diagnostic information provided by independently generalizable scores on each dimension is traded-off with the parsimony in information and reduced cost of a design to obtain generalizable composite scores.

The final design uses Equations 4.4 and 4.6 to derive an optimal design and optimal weights for the composite score. As seen in Table 4.2, to achieve a generalizability coefficient of 0.90 for the optimal composite, only 37 respondents are required to rate 5 chains on 2 service quality items and 1 product quality item. The weights are 0.998 for service quality and 0.057 for product quality. The weights imply that the chains are better discriminated on service quality than product quality. The reasons could be two-fold. The chains might truly differ more on service quality than on product quality; this is a reflection of greater variability in service offerings and higher standardization in the quality of products across retail chains. Secondly, the result might also be due to the quality of the measures, and therefore the costs of improving the measures. The univariate analysis suggested that the measurement of service quality was of better quality than that of product quality. Thus, the weights are also indicative of the quality of measures. Which of the two reasons is more dominant cannot be ascertained, primarily because it is impossible to measure "true" service or product quality.

4.5.2.2 Problem 2: Comparison of Respondents

The issue of comparison of respondents on their service quality and product quality perceptions arises in the context of segmentation. If the intent of a retail chain is to develop segments of customers, then it is important for the chain to obtain generalizable scores for comparing respondents on each dimension. The object of measurement for this problem is respondents, and the facets of generalization are retail sectors, retail chains, and items.

The optimal designs for this problem are in Table 4.3. The first case is a comparison of service quality perceptions only. Because of the high variance component associated with the interaction of respondents and chains, *42* chains are needed to obtain a

generalizability coefficient of 0.90 for comparing service quality perceptions across respondents. The number of items required are 5, leading to a total cost of $3510. The requirements for product quality are even more stringent; respondents have to rate the product quality of 47 chains from within each of 5 retail sectors on 6 different items. The cost of this design is $ 22710. A design to achieve generalizable scores for both dimensions requires respondents to rate the service quality of 42 chains on 5 service quality items and product quality of 47 chains within each of 5 retail sectors on 6 product quality items. The total cost is $ 26070.

Insert Table 4.3 about here

Table 4.3 also shows the optimal designs for an equally weighted composite and an optimally weighted composite. The equally weighted composite is expected to have a generalizability coefficient of 0.90 when respondents rate 39 retail chains from 5 retail sectors on 3 service quality items and 3 product quality items. The cost of such a design is $ 18870, lower than the cost of the design to achieve independently generalizable scores on each dimension. An even cheaper design is obtained when the composite is weighted optimally. A generalizability coefficient of 0.90 is expected when respondents rate 14 chains from 1 retail sector on 4 service quality and 4 product quality items. The optimal composite has weights of 0.853 for service quality and -0.521 for product quality. The cost of this design is $1942, substantially lower than any other design that measures both service and product quality.

Note that the composite has a negative weight, which is difficult to interpret. Moreover, the composite becomes very difficult to interpret as a meaningful construct. We take the view that although such a composite may lead to a lower cost design, it is difficult to interpret and therefore should not be used. Such a view is consistent with the opinion expressed elsewhere about composites with negative weights (Peter, Churchill,

and Brown 1993). We also view the negative weight as an indication that the comparison of the levels of the object of measurement, in this case respondents, will be better if one of the variables is left out of the composite. Thus, it essentially implies that the product quality, as measured by this procedure, does not help discriminate between respondents. An indication of this is provided by the optimal composite when we restricted the weights to be non-negative. The last design in Table 4.3 provides the result that the optimal composite should have a weight of 1 for service quality and 0 for product quality. The number of items was restricted to be positive, and therefore the number of product quality items is 1; the cost of this design reflects this additional item. Otherwise, the design and cost is identical to the first design for measuring service quality only. However, we would like to emphasize that the fact that a dimension such as product quality is given a weight of zero does not imply that it should be ignored. It simply implies that there are few differences across respondents on product quality perceptions.

An alternative approach to interpret negative weights is to examine the conditions under which a composite with negative weights will be more generalizable than a composite with positive weights. An investigation of this perspective may also shed some light on when the reliability of a difference score can be expected to be greater than the reliability of a simple average score across dimensions. To conduct a preliminary investigation of this issue, we attempted to find the conditions that lead to the reliability of difference score being greater than the reliability of a "sum" score (an average score). The investigation revealed that the conditions depend on the sign of the covariance components associated with the object of measurement and relative error. Basically, there are two distinct situations, outlined next.

Situation 1 occurs when the covariance components for the object of measurement and the relative error are both positive or when the object covariance is negative and error covariance is positive. Then the reliability of a difference score is greater than the reliability of sum score when,

$$(4.9) \qquad \frac{\sigma^2_{1,\,\text{object}} + \sigma^2_{2,\,\text{object}}}{\sigma^2_{1,\,\text{relative error}} + \sigma^2_{2,\,\text{relative error}}} > \frac{\text{cov}(1,\,2)_{\text{object}}}{\text{cov}(1,\,2)_{\text{relative error}}},$$

where $\hat{\sigma}^2_{i,\,\text{object}}$ represents the variance component for the i^{th} dimension associated with the object of measurement and $\text{cov}(1, 2)_{\text{object}}$ represents the covariance component (between dimensions 1 and 2) associated with the object of measurement. Similar interpretations can be made for the relative error variance-covariance components in Equation 4.9.

Situation 2 occurs when both covariances are negative, or when the object covariance is positive but the error covariance is negative. In this situation, the condition in Equation 4.9 is reversed, for the reliability of a difference composite to be greater than the reliability of a sum composite. To interpret this finding, we set the variance components for both dimensions for both the object and relative error to be equal to each other. Then, for *both situations*, if the covariance component associated with the relative error is greater than the covariance component associated with the object of measurement, then the reliability of a difference score will be greater than the reliability of a sum score. This preliminary investigation has suggested the algebraic conditions that may lead to a difference composite being more reliable than a sum composite. More work is required on the substantive interpretation of this finding, specially because of the widespread prevalence of difference score composites in the marketing literature.

4.6 Discussion and Conclusions

The empirical illustration on retail quality was intended to demonstrate the method in a context that might be intuitively appealing. The resultant designs from problem 1, comparison of retail chains, shows that there are different ways to use the information from the measurement of a multidimensional construct such as retail quality. If the purpose is to obtain generalizable scores of service quality only, then the design would be inexpensive relative to a purpose of obtaining generalizable scores of product quality. Obtaining independently generalizable scores for each dimension implies selecting the higher number of levels for those facets that are common to both

dimensions. In our illustration, respondents were a common facet across both dimensions; thus, 79 respondents were chosen for the combined study, even though the comparison of chains on service quality required only 38 respondents.

The optimal design for obtaining a generalizable equally weighted composite was shown to cost about half of what it would cost to obtain independently generalizable scores for each dimension. Thus, the trade-off between parsimony and diagnosticity was demonstrated by this illustration. In addition, we demonstrated how optimal numbers of levels can be selected for minimizing cost, while choosing the optimal weights for maximizing generalizability. A higher weight for a dimension indicates a combination of greater "true" differences on the dimensions across the object of measurement and/or "better" quality of measurement on the dimensions for the specific object of measurement.

There is a potential for a misleading managerial interpretation of the "optimal weights" as being indicative of the importances of the dimensions. Such a situation could arise if the manager decided that a dimension is not important because of a low weight produced by this analysis. This in turn could lead the manager to ignore the dimension, perhaps at the cost of ignoring the potential to make changes on the dimension that might lead to benefits such as higher profits. It should be made clear at this stage that we do not recommend ignoring a dimension in case of a low weight; we emphasize that a low weight implies either a situation of no differences across the levels of an object of measurement and/or greater error in the measurement of the object.

At the outset, we set out to answer two important questions. First, we asked if the reliability assessment methods for multidimensional constructs are any different from those for unidimensional constructs. We can now say that the reliability assessment method is not theoretically different. However, the method to assess reliability depends on whether or not there is an interest in forming a composite. The focus of the paper has been on composites, primarily because the formation of a composite score from

individual dimensions implies that the calculation of the generalizability coefficient include the vector of weights used to form the composite. The second question we asked was whether optimal designs for future decision studies depend on multidimensionality of the construct. From the results obtained in the study of retail quality, we can now say that the optimal design depends on two factors, both of which are related to the use of the information obtained from a decision study. Is the interest in obtaining independently generalizable scores for each dimension or is the interest in obtaining a generalizable composite of all dimensions? Independently generalizable scores for each dimension imply a more expensive design, although there are advantages of diagnosticity. A generalizable composite implies a less expensive design, but at the potential cost of unreliable diagnostic information. Thus, it is clearly a trade-off between reliability of the diagnosticity of information, parsimony, and cost of design. Secondly, the weights given to each dimension of a composite determine the optimal design. Thus, the composite can be improved to reflect a better discrimination between the levels of the object of measurement. Such a design is clearly the cheapest design in terms of monetary costs, because there are trade-offs made in terms of obtaining much less reliable information on each dimension and the loss of interpretability, as we saw in the case of the optimal composite for the second problem of comparing respondent perceptions. A natural application of this framework is the assessment of the generalizability of a composite formed with certain weights determined by theory. It is entirely possible that the composite has very poor generalizability; thus, weights determined by theory do not necessarily lead to good empirical measures. The optimization framework can also suggest ways to improve the design of a study that uses such a composite, so that the composite is made more generalizable.

Among the limitations of the methodology are the possibility of negative definite variance-covariance components matrices and the potentially difficult interpretation of negative weights. Future research can also explore substantive interpretation of the

conditions that lead to negative weights. Preliminary investigation suggests some conditions that lead to higher or lower multivariate generalizability coefficients, without altering the univariate generalizability coefficients. A high covariance component associated with the object of measurement, with low covariance components associated with all effects that are part of the relative error leads to a higher multivariate generalizability coefficient than for a situation with low covariance components for the effect associated with the object and high covariance components for the effects associated with the relative error. In addition, in the two dimensional case, the multivariate generalizability coefficient for the composite will be higher than each univariate generalizability coefficient if and only if the ratio of the object covariance component and the sum of the relative error and object covariance components is higher than each univariate generalizability coefficient. We call this investigation preliminary because it is based on restrictive assumptions, and is limited to the two-dimensional case explored here. For example, it assumes equivalence of the numbers of levels for each facet of generalization across each dimension. This is not a necessary condition in an actual decision study design.

To summarize, we presented a methodological framework to evaluate the reliability of multi-dimensional measurement, commonly conducted in the marketing literature and practice. In addition, we also presented an optimization method that extends Finn and Kayandé's (1997a) framework to multidimensional measurement. The contribution of the article is to present different methods to assess reliability under different uses of the information, and to show that the optimal measurement design depend on the way in which the information is intended to be used.

Appendix 4.1

Directions: The following statements ask how you feel about the service and products provided by some department store chains, grocery store chains, and fast-food chains. Please indicate the extent of your agreement with each statement about each chain. Circle a '10' if you very strongly agree, and circle a '0' if you very strongly disagree. If your feelings lie between these two extremes, circle a number in between '10' and '0' that best shows your level of agreement. There are no right or wrong answers- we are interested in your views of the service provided by the chains.

The following statements are about Eaton's department store chain.

1. Eaton's stores are visually attractive.
2. Eaton's employees appear neat and tidy.
3. Eaton's promotional materials are visually appealing.
4. Eaton's employees give you prompt service.
5. Eaton's employees are always willing to help you.
6. Eaton's employees are never too busy to respond to your requests.
7. Eaton's employees give you personal attention.
8. Eaton's employees have your best interests at heart.
9. Eaton's employees understand your specific needs.
10. The products available at Eaton's are of high quality.
11. Eaton's has all the items I want to buy at a departmental store.
12. Eaton's has a good selection of quality products.

Note: Each statement was accompanied by an 11-point scale anchored at the end-points by the labels "Very Strongly Disagree" (= 0) and "Very Strongly Agree (= 10). The intermediate scale points were not labeled. Also, the statements were not numbered.

Appendix 4.2

Cost Function used in the Optimization

Extending Finn and Kayandé's (1997a) cost function to the multivariate case, cost C of a multivariate survey design is given by,

$$C = f(c_0, \tilde{C}, \tilde{N}, n_{chains}) = c_0 + c_1 N_{r\,(max)} + c_2 (n_{i(sq)} + n_{i(pq)})\, n_c\, n_s + c_3 (n_{i(sq)}\, n_{r(sq)} + n_{i(pq)}\, n_{r(pq)})\, n_c n_s$$

where,

1. c_0 is the fixed cost of the study.
2. c_1 is the unit cost of selecting and communicating with a respondent.
3. c_2 is the unit cost of an additional item i when designing and formatting the data collection instrument.
4. c_3 is the incremental cost of a lengthening of the study with an additional item on the data collection cost for each respondent.
5. c_0 = \$0, c_1 = \$5, c_2 = \$10, c_3 = \$0.20 for both problems 1 and 2.
6. $n_{r(sq)}$ is the number of respondents who respond to the service quality items, $n_{i(sq)}$ is the number of service quality items.
7. $n_{r(pq)}$ is the number of respondents who respond to the product quality items, $n_{i(pq)}$ is the number of product quality items.
8. $N_{r(max)}$ is the greater of $n_{r(sq)}$ and $n_{r(pq)}$. When both dimensions are being measured simultaneously, every respondent responds to items on both dimensions. Thus, the set of respondents is common across both dimensions. In addition, in calculations of cost (for designs where both dimensions are measured simultaneously) in Tables 4.2 and 4.3, we replace $n_{r(sq)}$ and $n_{r(pq)}$ with $N_{r(max)}$.
9. For any problem where the design is being optimized for one dimension only, all terms that involve the other dimension are equal to zero.
10. n_c is the number of retail chains, and n_s is the number of retail sectors.
11. For problem 1, n_c is equal to 5 and n_s is equal to 1. For problem 2, $N_{r(max)}$ is equal to 30.

Table 4.1: Estimated Variance-Covariance Components Matrices

Source of Variation	Service Quality	Product Quality
Retail Sector	0.000[a]	-0.008
	-0.008	1.048
Retail Chain	0.623	0.395
	0.395	0.457
Respondents	1.003	0.804
	0.804	1.123
Items	0.076	0.000
	0.000	0.000[b]
Sector X Respondent	0.027	0.211
	0.211	0.186
Sector X Items	0.011	0.000
	0.000	0.072
Chains X Respondents	1.698	0.898
	0.898	1.970
Chain X Items	0.029	0.000
	0.000	0.218
Respondents X Items	0.164	0.000
	0.000	0.222
Sector X Respondents X Items	0.034	0.000
	0.000	0.223
Error	0.681	0.000
	0.000	1.094
Generalizability Coefficient for Comparing Retail Chains	0.205	0.122
Generalizability Coefficient for Comparing Respondents	0.278	0.233

[a]Truncated to zero from -0.007
[b]Truncated to zero from -0.0003

Table 4.2: Comparing 5 Retail Chains on two dimensions independently, both dimensions simultaneously, and a composite of two dimensions

	No. of Respondents	No. of SQ[a] Items	No of PQ Items	Cost ($)	G-coeff. for SQ	G-coeff. for PQ	G-coeff. for composite
1. Only Service Quality (SQ)	38	2	-	366	0.901	-	-
2. Only Product Quality (PQ)	79	-	9	1556	-	0.900	-
3. Simultaneous SQ and PQ (No composite)	79	1	9	1685	0.913	0.900	-
4. Simple Composite *(0.707SQ + 0.707PQ)*	62	1	3	758	0.902	0.805	0.90
5. Optimal Composite *(0.998SQ + 0.057PQ)*	37	2	1	446	0.900	0.603	0.90

Table 4.3: Comparing 30 respondents on two constructs independently, both constructs simultaneously, and a composite of two constructs

	No. of Retail Sectors	No. of Retail Chains	No. of SQ Items	No of PQ Items	Cost ($)	G-coeff. for SQ	G-coeff. for PQ	G-coeff. for composite
1. Only Service Quality	1	42	5	-	3510	0.900	-	-
2. Only Product Quality	5	47	-	6	22710	-	0.900	-
3. Simultaneous SQ and PQ (No composite)	1, 5[a]	42, 47[a]	5	6	26070	0.900	0.900	
4. Simple Composite *(0.707SQ + 0.707PQ)*	5	39	3	3	18870	0.903	0.863	0.90
5. Optimal Composite *(0.853SQ - 0.521PQ)*	1	14	4	4	1942	0.826	0.710	0.90
6. Optimal Composite (+ve weights only) *(1.000SQ + 0.000PQ)*	1	42	5	1	4182	0.900	0.615	0.90

[a] Sample size requirements for service quality and product quality respectively.

Bibliography

Armor, David J. (1974), "Theta Reliability and Factor Scaling," in H.L. Costner (ed) Sociological Methodology, San Francisco: Jossey-Bass, 17-50.

Behrman, Douglas, N. and William D. Perreault, Jr. (1982), "Measuring the Performance of Industrial Salespeople," Journal of Business Research, 10 (September), 355-70.

Bienstock, Carol C., John T. Mentzer, and Monroe Murphy Bird (1997), "Measuring Physical Distribution Service Quality," Journal of Academy of Marketing Science, 25 (1), 31-44.

Bock, R.D. (1963), "Multivariate Analysis of Variance of Repeated Measurements," in Problems of Measuring Change, C.W. Harris (ed.), Madison: University of Wisconsin Press.

Bock, R.D. (1966), "Contributions of Multivariate Experimental Design to Educational Research," in Handbook of Multivariate Experimental Psychology, R.B. Cattell (ed.), Chicago: Rand McNally.

Conger, A.J. and R. Lipshitz (1974), "Measures of Reliability for Profiles and Test Batteries," Psychometrika, 38, 411-427.

Cronbach, Lee J., Goldine C. Gleser, Harinder Nanda, and Nageswari Rajaratnam (1972) The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles. New York: John Wiley & Sons.

Dabholkar, Pratibha, Dayle I. Thorpe, and Joseph O. Rentz (1996), "A Measure of Service Quality for Retail Stores," Journal of Academy of Marketing Science, 24 (1), 3-16.

Finn, Adam and Ujwal Kayandé (1997a), "Reliability Assessment and Optimization of Marketing Measurement," Journal of Marketing Research, 34 (may), 262-75.

Finn, Adam and Ujwal Kayandé (1997b), "A More General Paradigm for Scale Development in Marketing," Working Paper, University of Alberta.

Finn, Adam and Ujwal Kayandé (1997c), "Unmasking a Phantom: A Psychometric Assessment of Mystery Shopping," Working Paper, University of Alberta.

Heise, David R. and George W. Bohrnstedt (1970), "Validity, Invalidity and Reliability," in E.F. Borgatta and G.W. Bohrnstedt (eds.) Sociological Methodology. San Francisco: Jossey-Bass 104-129.

Joe, George W. and J. Arthur Woodward (1976), "Some Developments in Multivariate Generalizability," Psychometrika, 41(2) 205-217.

MacKenzie, Scott, Philip M. Podsakoff, and Richard Fetter (1993), "The Impact of Organizational Citizenship Behavior on Evaluations of Salesperson Performance," Journal of Marketing, 57 (1), 70-80.

Nunnally, Jum C. (1978), Psychometric Theory. Second Edition. New York: McGraw-Hill.

Parasuraman, A., Leonard L. Berry, and Valarie A. Zeithaml (1991), "Refinement and Reassessment of the SERVQUAL Scale," Journal of Retailing, 67 (Winter), 420-50.

Parasuraman, A., Valarie A. Zeithaml, and Leonard L. Berry(1988), "SERVQUAL: A Multiple-Item Scale for Measuring Consumer Perceptions of Service Quality," Journal of Retailing, 64 (Spring), 12-40.

Peter, J. Paul (1979), "Reliability: A Review of Psychometric Basics and Recent Marketing Practices," Journal of Marketing Research, 16, (February), 6-17.

Peter, J. Paul, Gilbert A. Churchill, Jr., and Tom J. Brown (1993), "Caution in the Use of Difference Scores in Consumer Research," Journal of Consumer Research, 19 (4), 655-62.

Rentz, Joseph O. (1987), "Generalizability Theory: A Comprehensive Method for Assessing and Improving the Dependability of Marketing Measures," Journal of Marketing Research, 24 (February), 19-28.

_____ (1988), "An Exploratory Study of the Generalizability of Selected Marketing Measures," Journal of the Academy of Marketing Science, 16 (Spring), 141-150.

Searle, Shayle R., George Casella, and Charles E. McCulloch (1992), Variance Components. New York: John Wiley and Sons.

_____ and T.R. Rounsaville (1974), "A Note on Estimating Covariance Components," American Statistician, 28, 67-68.

Weitz, Barton (1978), "The Relationship Between Salesperson Performance and Understanding of Customer Decision Making," Journal of Marketing Research, 15 (November), 501-16.

Webb, Noreen M. and Richard J. Shavelson (1981), "Multivariate generalizability of general educational development ratings," Journal of Educational Measurement, 18(1) 13-22.

Woodward, J Arthur. and George W. Joe (1973), "Maximizing the Coefficient of Generalizability in Multi-facet Decision Studies," Psychometrika, 38(2) 173-181.

Chapter 5

General Discussion and Conclusions

The central theme of this thesis is that the measurement of a characteristic of an object is influenced by the conditions under which the characteristic is measured. Thus, information gathered by a measurement procedure is dependent on the state of the object at the time of measurement. Additionally, informational dependencies are created when only a subset of conditions that affect the measurement of the object's characteristic are explicitly recognized in a measurement study. In the face of such potential informational dependencies, the generalizability of the information gathered is restricted to the domain of inquiry for that measurement study. The potential of informational dependencies exists not only in the multi-item survey measurement procedures examined in this thesis, but also across other procedures such as the analysis of scanner panel data or observational methods. For example, any substantive conclusion about the effect of sales promotions on a performance criterion, such as sales or market shares, has to take into account the potential for the variation in the results across product categories. That such studies are conducted across different product categories reflects the possibility of a variation in results across categories. A meta-analysis of these results can then be used to determine if such variation is systematic, and the magnitude of the systematic variation (Farley and Lehmann 1986). A meta-analysis often allows the researcher to find interesting informational dependencies that otherwise were not identified in the literature. Farley and Lehmann (1986) imply then that a meta-analysis primarily investigates the empirical *generalizability* of substantive research conclusions.

This thesis *generalizes* the study of generalizability by suggesting that such questions can be asked *a priori*, rather than post-hoc as is the case with a meta-analysis. It is not my argument that every study that is conducted must include all the conditions that might affect the measurement. To suggest thus would clearly reflect a lack of research experience. Research domains have to be limited because it is not possible to examine the

variation due to every condition in the universe in a single study. However, it is certainly possible to recognize not only the limitations of the domain of inquiry, but also the potential methods to generalize the substantive results. Generalizability theory (Cronbach et al. 1972) offers a powerful methodology to examine variation due to different conditions of measurement.

Additionally, this thesis argues that it is of paramount interest for marketers to recognize the purpose of measurement, whether a study has academic or managerial purposes. High reliability when one facet is the object of measurement does not ensure high reliability when other facets are the object of measurement. That the same data can provide "good" information for one purpose and "poor" information for another purpose, is unfortunately not recognized in the marketing literature. Generalizability theory ". . . enables you to *ask* your questions better; what is most significant for you cannot be supplied from the outside" (Cronbach et al. 1972, pg. 199, italics added). This thesis is a demonstration of the theory within the domain of marketing measurement, and the ability of the theory to improve the quality and efficiency of marketing measurement. To my knowledge, it is the first complete application of the theory within marketing to a substantive marketing measurement issue.

This thesis uses the previous arguments to suggest and demonstrate that the use of classical test theory methods in marketing measurement evaluation is flawed to the extent that, (1) the measurement is affected by different conditions under which the characteristic is measured, (2) the measurement studies have purposes other than scaling respondents, (3) future measurement studies can be efficiently designed on the basis of past information, and (4) the characteristic being measured may in some cases be multidimensional.

The first paper of the thesis provided a demonstration of the methodology to the measurement of service quality, treating the dimensions of service quality as different levels of a single factor. The service quality application illustrates a method of designing

measurement instruments so that repeated measurement studies can be optimal in terms of data collection requirements, psychometric standards, and the cost of measurement. The optimization is primarily based on the identification of the object of measurement, which is dependent on the purpose of measurement, and knowledge of the relative sizes of the variance components. The paper also demonstrates a simple integer programming optimization approach, where the cost of measurement is minimized subject to a constraint of a pre-specified generalizability. This summary of the first paper suggests that the primary contribution of the first paper is in demonstrating the methodology on the measurement of a substantively important marketing issue, and the possibility of designing efficient measurement studies.

The first paper also points to the importance of designing applied measurement studies using both statistical principles and substantive insights. For example, the possibility that a certain measurement task requires a large number of items could be construed as a problem in traditional survey research because of fatigue and non-response problems. However, in the generalizability theory formulation, the use of mystery shoppers would overcome this problem. The ability to compare levels of an object with an "eye" for detail might also be an argument to use mystery shoppers over traditional survey respondents. Traditional survey respondents essentially report on their natural experiences, which have sometimes been acquired several months before the survey response. The responses from mystery shoppers are more current, because they respond immediately after the experience. Thus, this provides an additional contribution of the extant work.

The methodology flowing from generalizability theory is statistical in nature and much more complex than the simple methods used in applying classical test theory methods. This has led to a neglect of the theory in marketing as well as a continuing dependence on classical theory methods. This thesis provides the first full application of the theory in marketing, and as a result identifies complexities such as unbalanced data

that do not normally occur in applications of generalizability theory in other disciplines (e.g. education and psychology). Such complexities bring into play the need to recognize the availability of different estimation methods, the focus of the second paper of the thesis. The paper estimates confidence intervals for variance components, an issue considered extremely important by the original proponents of generalizability theory, but rarely investigated in the measurement application literature. The second paper of the thesis showed that optimality of measurement designs depends on the different estimation methods, but not in any systematic way. The paper also alerted researchers to the possibility that the intervals on variance components were such that it is difficult to place much confidence at the current time in the optimality of designs constructed using methods suggested in the first paper of the thesis. This is so *not* because of the inadequacies of the methodological framework from the first paper, but because of the wide intervals on variance components estimated from a small number of levels of facets. Thus, the paper clearly implies that future research is required on how to better apply the framework developed in the first paper.

Finally, the thesis recognizes that an object's characteristics can be multidimensional. Salesperson performance measurement (Behrman and Perreault 1982, Bush et al. 1990), salesperson job satisfaction (Churchill, Ford, and Walker 1974), service quality (Parasuraman, Zeithaml, and Berry 1988), market orientation (Narver and Slater 1990), emotional responses to advertisements (Holbrook and Batra 1987), customer market power (Butaney and Wortzel 1988), organizational buying power (Kohli 1989), channel member satisfaction (Ruekert and Churchill 1984) are a few examples of multidimensional characteristics that abound in marketing. The assessment of the generalizability of multidimensional measurement requires a multivariate extension of generalizability theory. The final paper of the thesis suggests the appropriate methodology and provides a demonstration of the methodology on the measurement of the quality of a retail outlet. Multivariate generalizability theory provides an additional

insight of interest into multidimensional concepts. In the paper, I showed that it is possible to determine optimal weights for the dimensions to form a composite score. The composite score could be used to compare different levels of the object, or could be used as an explanation for some related phenomenon. For the comparison of levels of the object of measurement, the optimal weights represent the weighting of dimensions such that the generalizability of such a comparison is maximal. Although there is high possibility that an optimally weighted composite will better explain relationships with other characteristics of the levels of the object, I have not investigated this interesting possibility in this thesis.

As far as managerial contributions are concerned, the thesis demonstrates a methodology to efficiently design measurement studies of managerial interest. The first paper clearly demonstrates the importance of taking into account the *management problems* underlying the need for the measurement of service quality. The claim made in this thesis that the assessment of the "quality" of information is dependent on the managerial purpose implies that the methodology is flexible enough to accommodate diverse managerial purposes. The estimation methods, although highly statistical in nature, have much to do with the dependability of the estimates of variance components. The estimates of variance components are the key input to the optimization exercise and thus their dependability has to be of prime interest to managers who are potential users of this methodology. The methodological framework can be applied to several different measurement problems, including efficient new product concept testing, the assessment of customer service by mystery shopping, the assessment of quality of public transport, and the assessment of quality of store image measurement. Other areas of potential fruitful applications include conjoint analysis for new product assessments, assessment of advertising effectiveness, and brand equity.

Future research opportunities in this area of research are aplenty. Most of the opportunities flow from the limitations of the work in this thesis, which I have outlined below.

Negative variance estimates (or negative definite variance-covariance components matrices) are a potential problem in most variance components models estimated with methods that do not put constraints on parameter estimates. In such cases, Cronbach et al. (1972) and Brennan (1983) have recommended setting the negative variance to zero. They differ on the procedures to adjust the other variance components after setting a negative variance estimate to zero. Cronbach et al. (1972) suggest that the other variance components should not be adjusted in order to maintain unbiasedness of estimates. Brennan (1983) suggests recalculating the variance estimates adjusting for the negative variance estimates set to zero. In the second paper of the thesis, I suggested the use of maximum likelihood methods for preventing negative variance estimates, although this results in biased estimates. Alternatively, if some distributional assumptions are made and computational complexity is accepted, Bayesian procedures prevent the problem of negative variance estimates (Brennan 1983). Bayesian variance components models are now in use and several researchers in genetics have explored Bayesian methods to estimate variance components. It is expected that Bayesian methods would be used for estimation in the near future because "generalizability theory . . . is Bayesian in everything but a formal sense" (Novick 1976).

Generalizability theory also assumes that variance components remain stable from generalizability to decision studies. This assumption is critical for the optimization of measurement to remain valid. We speculate that only major changes in a market will cause instability in variance components. But it is also possible to verify the assumptions made in the optimization exercise by reexamining the variance components in the applied study.

An important assumption is made in this thesis with respect to the optimization. The optimization assumes that there will be no missing data in the decision studies. This strong assumption is likely to be violated often in most methods of data collection. One way to relax this assumption when generating optimal designs might be to assume that the distribution of missing cells across different facets in the generalizability study will hold for the decision study, and then generate optimal designs taking it into account. Thus, the number of levels of an individual facet will increase in proportion to the expectation of missing data on that facet.

Statistical sampling is one way to generate optimal designs. Statistical design issues, such as whether a factor(s) is crossed or nested, and fixed or random, also determine the optimal design. Moreover, the efficiency of a measurement design will depend on the form of data collection and the format of the stimulus material used in the study if the layout of a questionnaire affects the variance components in a systematic manner. For example, instead of putting all items for a chain in one block as in the layout used for this thesis, the research study could have used a questionnaire with a block of chains for each item. If respondents respond to the stimulus by first setting a mean for a block and then distributing ratings around the mean, there could be predictable changes in the variance components due to items and chains. If the purpose of measurement is to differentiate between chains, the optimal format to be used may not be the same as if the purpose is to differentiate between items. Thus, the layout of the questionnaire could be used to make a scale more generalizable. Similarly, alternative data collection methods will not only have different functional forms for the cost, they might be expected to generate different variance components, resulting in quite different optimal measurement designs. What remains to be explored is the extent to which such methods systematically influence the generalizability of a scale.

The results from the second paper of the thesis imply that the optimality of designs developed from point estimates of variance components may be questionable

because of the width of the confidence intervals around the estimates. Further research should investigate the development of optimal designs using information on the confidence intervals around the estimates of variance components. Thus, the optimization conditions would be expanded to include an *a priori* restriction on width of the interval around the generalizability coefficient. As a first step, a simulation should provide an idea of the impact of this additional condition.

Bibliography

Behrman, Douglas and William D. Perreault, Jr. (1982), "Measuring the Performance of Industrial Salespersons," Journal of Business Research, 10, 355-370.

Brennan, Robert (1983), Elements of Generalizability Theory. Iowa City, Iowa: ACT Publications.

Bush, Robert P., Alan J. Bush, David J. Ortinau, and Joseph F. Hair (1990), "Developing a Behavior-Based Scale to Assess Retail Salesperson Performance," Journal of Retailing, 66 (Spring), 119-129.

Butaney, Gul and Lawrence H. Wortzel (1988), "Distributor Power Versus Manufacturer Power: The Customer Role," Journal of Marketing, 52 (January), 52-63.

Churchill, Gilbert A., Jr., Neil M. Ford, and Orville C. Walker (1974), "Measuring the Job Satisfaction of Industrial Salesmen," Journal of Marketing Research, 11 (August), 254-260.

Cronbach, Lee J., Goldine C. Gleser, Harinder Nanda, and Nageswari Rajaratnam (1972) The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles. New York: John Wiley & Sons.

Farley, John U. and Donald R. Lehmann (1986), Meta-Analysis in Marketing: Generalization of Response Models. Lexington, MA: Lexington Books.

Holbrook, Morris B. and Rajeev Batra (1987), "Towards a Standardized Emotional Profile Useful in Measuring Responses to the Nonverbal Components of Advertising," in Nonverbal Communications in Advertising, Sidney Hecker and David W. Stewart (Eds.), Lexington, MA: D.C. Heath, 95-109.

Kohli, Ajay K. (1989), "Determinants of Influence in Organizational Buying: A Contingency Approach," Journal of Marketing, 53 (July), 50-65.

Narver, John C. and Stanley F. Slater (1990), "The Effect of Market Orientation on Business Profitability," Journal of Marketing, 54 (October), 20-35.

Novick, M.R. (1976), "Bayesian Methods in Educational Testing: A Third Survey," in Advances in Psychological and Educational Testing, D.M.N. de Gruijter and L.J.T. van der Kamp (Eds.), NewYork: Wiley.

Parasuraman, A., Valarie A. Zeithaml, and Leonard L. Berry (1988), "SERVQUAL: A Multiple-Item Scale for Measuring Consumer Perceptions of Service Quality," Journal of Retailing, 64 (Spring), 12-40.

Ruekert, Robert W. and Gilbert A. Churchill, Jr. (1984), "Reliability and Validity of Alternative Measures of Channel Member Satisfaction," Journal of Marketing Research, 21 (May), 226-233.

Appendix I

Statistical Model Underlying Measurement in Classical Test Theory and Generalizability Theory

Classical Test Theory: *Underlying Statistical Model and the Reliability Coefficient*

Assuming individuals are measured on some characteristic X, the random variable X with mean μ_x and variance σ^2_x, can be represented as the sum of a true score random variable T with mean μ_t and variance σ^2_t and error score random variable E with mean μ_e and variance σ^2_e. Further, it can be shown that μ_e is equal to zero (Traub 1994, pg. 31). This representation is derived from the fundamental equation of classical test theory (which is a tautology and therefore, is not falsifiable),

(A1) $$x_p = \tau_p + e_p$$

where, x_p is the p^{th} individual's observed score, τ_p is p^{th} individual's true score, and, e_p is the error in measuring the p^{th} individual's true score.

Over more than one individual, Equation A1 can be written as,

(A2) $$X = T + E$$

Given the definition of variances, and the independence of T and E (Traub 1994, pg. 32), i.e., the covariance between T and E is zero, the variance of the observed score X can be written as,

(A3) $$\sigma^2_x = \sigma^2_t + \sigma^2_e$$

The reliability coefficient is defined as the intraclass correlation coefficient and is indicative of the ability of the measure to ". . . consistently rank-order and maintain the distance between subjects (upto a linear transformation)" (Peter 1979). It is expressed as the proportion of observed variance that is due to the true score variance,

(A4) $$\rho_x^2 = \frac{\sigma_t^2}{\sigma_x^2}$$

If the measure is capable of discriminating between individuals, then the reliability will be relatively high. If all individuals obtain similar scores, then the true score variance will be zero and therefore, reliability will be zero. Reliability can be assessed in many different ways including test-retest reliability, internal consistency, and alternative forms reliability; however, the fundamental basis for the estimation of reliability is Equation A4. The most popular estimate of reliability (internal consistency) in the marketing literature is coefficient α or Cronbach's α (Cronbach 1951), which is expressed as,

(A5) $$\alpha = \frac{n}{(n-1)}\left[1 - \frac{\sum_{i=1}^{n} \sigma_i^2}{\sigma_t^2}\right]$$

where,

σ_i^2 is the variance of each item i, i = 1, 2, 3, ... n

σ_t^2 is the total variance of the n-item scale

As is obvious in Equation A5, this formula is strictly for multi-item scales. This formula provides an estimate of the reliability coefficient, as defined in Equation A4, only when the items are parallel (the items have equal means and variances). However, such parallel tests/items are difficult to construct in most measurement contexts, in which case coefficient α provides an estimate that is the lower bound of the reliability coefficient.

A close examination of either Equation A4 or Equation A5 shows that error is treated as essentially undifferentiated in classical test theory. In Equation A4, error is assumed to be inherently random. Equation A5, on the other hand, assumes that error in measurement comes from the items not measuring identical constructs and therefore, increasing the number of items decreases error to the extent that the additional items are highly positively correlated. In either case, error originates from a single source and the object of measurement is always the individual. The model can be extended to include any object of measurement, however the assumption of single-faceted nature of error does not change.

Generalizability Theory: *Variance Components Models and Generalizability Coefficient*

As opposed to the limited view of classical test theory, generalizability theory (Cronbach et al. 1972) provides a multi-faceted view of measurement, where variation in measurement can arise from multiple controllable sources. Following the Fisherian logic of experimental design, Cronbach and his colleagues suggested expressing measurement in terms of random effects models. Thus, the "error formerly seen as amorphous is now attributed to multiple sources, and a suitable experiment can estimate how much variation arises from each controllable source" (Cronbach et al. 1972). Cronbach, Rajaratnam, and Gleser (1963) and Gleser, Cronbach, and Rajaratnam (1965) presented the theoretical foundations of generalizability theory. Thereafter, several studies in educational psychology have applied and extended generalizability theory. Prominent among the reviews of the literature on generalizability theory are Brennan (1983) and Shavelson and Webb (1981, 1991). The focus of this explanation is on the underlying statistical model, which is not explained elsewhere in the thesis.

Consider a simple problem of the measurement of service quality provided by grocery chains. Such a measurement problem arises when the management of a grocery chain is interested in comparing its service quality to that of its competitors. A marketing researcher will, in all probability, conduct a survey with multiple items (questions) that ask multiple respondents to evaluate the service quality provided by a number of grocery chains on a rating scale (ignoring, for the moment, the possibility that the items might be representative of substantively different aspects of service quality). Classical test theory methods will either assess the reliability of the multi-item scale for each individual grocery chain (treating them as distinct) or treat grocery chains as contributing no variance in the measure and therefore aggregate over all chains. However, familiar experimental design principles suggest at least 3 different systematic factors in this market research study, viz., grocery chains, respondents, and items. Now we proceed to partition the variation in ratings into these and associated sources of variation.

An observed rating for any grocery chain (i) by any respondent (j) on any item (k) is represented by X_{ijk}. The average rating for any grocery chain *over all respondents and all items* is denoted by μ_i. This average rating is taken to represent a *universe score* for a grocery chain because it is the *expected* value over all respondents and all questions in the universe of respondents and questions. This representation is based on the assumption of random sampling of items and respondents from the universe of items and respondents. Similarly, the expected rating given by a respondent j over all grocery chains and all items is μ_j and the expected rating on an item k over all respondents and grocery chains is denoted by μ_k. Similarly, μ_{ij} represents the average over all items, μ_{ik} represents the average over all respondents, and μ_{jk} represents the average over all grocery chains.

Finally, μ denotes the grand mean over all grocery chains, items, and respondents. Note that these values are not observable owing to the fact that they are population parameters.

An observed rating X_{ijk} can be expressed in terms of these parameters. The rating X_{ijk} given by respondent j to grocery chain i on item k can be partitioned as,

(A6)
$$
\begin{aligned}
X_{ijk} = {} & \mu && \text{(grand mean)} \\
& + \mu_i - \mu && \text{(grocery chain effect)} \\
& + \mu_j - \mu && \text{(respondent effect)} \\
& + \mu_k - \mu && \text{(item effect)} \\
& + \mu_{ij} - \mu_i - \mu_j + \mu && \text{(chain by respondent effect)} \\
& + \mu_{ik} - \mu_i - \mu_k + \mu && \text{(chain by item effect)} \\
& + \mu_{jk} - \mu_j - \mu_k + \mu && \text{(respondent by item effect)} \\
& + X_{ijk} - \mu_{ij} - \mu_{ik} - \mu_{jk} + \\
& \quad \mu_i + \mu_j + \mu_k - \mu && \text{(residual effect)}
\end{aligned}
$$

where every effect 's' has a distribution with mean zero and variance σ^2_s. All effects are assumed to be independent of each other. The model can be succinctly written as the random effects model,

(A7)
$$X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \varepsilon_{ijk}$$

where α, β, γ, etc. represent the random effects and ε represents the error term. Thus the variation in the ratings X_{ijk} can represented by,

(A8)
$$\sigma^2_x = \sigma^2_\alpha + \sigma^2_\beta + \sigma^2_\gamma + \sigma^2_{\alpha\beta} + \sigma^2_{\alpha\gamma} + \sigma^2_{\beta\gamma} + \sigma^2_\varepsilon$$

Equation A8 is the familiar variance components model (Searle, Casella, and McCulloch 1992). The estimated variance components are representative of the expected

variation in the population defined by the conditions observed in an experiment. All levels of each factor are assumed to be randomly sampled from an infinite population. The general model can be modified to include fixed effects and finite populations.

Generalizability Coefficient:

A research study is conducted for the purposes of theory development and/or gathering information for decision making. Studies conducted for the latter purpose are generally called applications or decision studies, whereas studies for the former purpose are appropriately called theoretical or generalizability studies. A large number of academic studies in marketing are conducted for the purpose of theory development; however, in recent years an increasingly large number of academic studies are being conducted for the purposes of measurement scale development. Such measurement scales are often recommended to practitioners for the purposes of marketing decision making. The ability of a measurement scale to reliably scale the characteristics of an object of measurement determines an important aspect of the quality of information that results from using the measurement scale. Cooil and Rust (1994) suggest that an intraclass correlation coefficient can be interpreted as an indicator of the loss expected from using the information resulting from the use of a measurement procedure. Thus, an intraclass correlation coefficient close to 1 indicates that the loss from using the measurement scale can be expected to be very small, whereas a coefficient close to 0 indicates the loss can be expected to be very large.

An intraclass correlation coefficient in the context of a variance components model representation of a measurement procedure has been defined within generalizability theory. The generalizability coefficient is analogous to and subsumes the

reliability coefficient. Before giving an expression for the coefficient, it is important to discuss the basis for the formula. Of the several sources of variance in a measure, there are some that constitute signals in the context of a specific application. For example, in the example of service quality of grocery chains, the variation attributable to the service quality provided by grocery chains might be of substantive interest to the decision maker. The service quality of grocery chains is then the characteristic being observed, grocery chains are the object of measurement, and different grocery chains constitute levels of the object. Then, the signal would be of better "quality" if the variation attributable to grocery chains is large relative to other sources of variance. The signal thus depends on the purpose of measurement, i.e., the use to which the information from the measurement will be put.

The error in the signal is of paramount interest in generalizability theory, for some of the error is controllable. A variance component of significance associated with the interaction of the object of measurement with any other source of variance implies that the scaling of the levels of the object of measurement (or the signal) is *dependent* on that source of variance. Thus, such measurement dependency implies that this variance component will constitute error in the measurement of the object of measurement. Another interpretation is that the scaling of the levels of an object of measurement should generalize over a larger universe, and therefore these sources of error are substantively defining the universe over which the information from the signal can be generalized. An option is to remove the source that contributes error, quite clearly implying that the universe of generalization becomes smaller. For example, if the service quality of grocery chains were assessed by a single respondent, and making the reasonable assumption that

there will be heterogeneity among respondents in their evaluations, it will not be possible to generalize the results from the single respondent to the universe of respondents. Thus, although the signal-noise ratio might be quite high, the universe of generalization will be highly limited and of little practical use.

Scaling grocery chains in terms of their service quality is an example of an application of a measurement procedure that results in information for the purposes of comparative decisions. In the context of the variance components model in Equation A8, the error in measurement for comparative decisions can be defined in terms of the variance components associated with the n-way interactions of grocery chains with respondents and items. The error in measuring the characteristic of interest is the sum of all variance components associated with the interaction of the object of measurement with all other sources of variance accounted for in the study. This error is called relative error variance in generalizability theory and is defined in the context of the grocery chain example to be,

(A9) $$\sigma^2_{\text{relative error}} = \frac{\sigma^2_{\psi}}{n_{\text{respondents}}} + \frac{\sigma^2_{\delta}}{n_{\text{items}}} + \frac{\sigma^2_{\varepsilon}}{n_{\text{respondents}} n_{\text{items}}}$$

where $n_{\text{respondents}}$ is the number of respondents, and n_{items} is the number of items.

The division of each component by the number of levels of the facet reflects the important property of a distribution of mean scores or effects that the variance of the distribution is equal to the variance of the individual elements divided by the respective sample sizes (Brennan 1983, Cronbach et al. 1972). The sample sizes are the number of levels for each facet.

These basic concepts allow for an expression of the generalizability coefficient (Cronbach et al. 1972), given in Equation A10:

(A10) $$E\rho^2 = \frac{\sigma^2_{\text{universe score}}}{\sigma^2_{\text{universe score}} + \sigma^2_{\text{relative error}}}$$

where $\sigma^2_{\text{universe score}}$ is the variation in effects associated with any object of measurement (analogue of the true score variance in classical test theory) and $\sigma^2_{\text{relative error}}$ is the sum of only those variance components that affect the scaling of the levels of the object of measurement. The notation $E\rho^2$ is meant to show that a generalizability coefficient is "approximately equal to the *expected value* . . . of the squared correlation between observed scores and universe scores" (Brennan 1983, p. 17, emphasis mine). The generalizability coefficient can be used in two important contexts. First, it can be used to examine the reliability of the information gathered in a generalizability study, if the information is intended to be used for decision making purposes. Second, it can be used to determine a sampling scheme for future applications such that the reliability of the information in the future is of a certain desired value. Thus, the generalizability coefficient can be used to recommend efficient designs which result in high reliability.

Bibliography

Brennan, Robert (1983), Elements of Generalizability Theory. Iowa City, Iowa: ACT Publications.

Cooil, Bruce and Roland T. Rust (1994), "Reliability and Expected Loss: A Unifying Principle," Psychometrika, 59 (June), 203-216.

Cronbach, Lee J. (1951), "Coefficient Alpha and the Internal Structure of Tests," Psychometrika, 16, 3, 297-334.

_____, Goldine C. Gleser, Harinder Nanda, and Nageswari Rajaratnam (1972) The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles. New York: John Wiley & Sons.

_____, Nageswari Rajaratnam and Goldine C. Gleser (1963), "Theory of Generalizability: A Liberalization of Reliability Theory," British Journal of Statistical Psychology, 16 (November), 137-163.

Gleser, Goldine C., Lee J. Cronbach, and Nageswari Rajaratnam (1965), "Generalizability of Scores Influenced by Multiple Sources of Variance," Psychometrika, 30, 395-418.

Peter, J. Paul (1979), "Reliability: A Review of Psychometric Basics and Recent Marketing Practices," Journal of Marketing Research, 16, (February), 6-17.

Searle, Shayle R., George Casella, And Charles E. McCulloch (1992), Variance Components. New York: John Wiley and Sons.

Shavelson, Richard J. and Noreen M. Webb (1981), "Generalizability Theory: 1973-1980," British Journal of Mathematical and Statistical Psychology, 34 (November), 133-166.

Shavelson, Richard J. and Noreen M. Webb (1991), Generalizability Theory: A Primer. Newbury Park, CA: SAGE Publications.

Traub, Ross E. (1994), Reliability for the Social Sciences. Thousand Oaks, CA: SAGE Publications.

Appendix II

STORES IN THE EDMONTON AREA

A SURVEY OF EDMONTON AREA RESIDENTS

Over recent years numerous retail organizations have gone bankrupt. Some critics have suggested half the existing organizations will not be around in the year 2000. However, little is known about what people like yourself think of the service provided by stores in the Edmonton area and how they are used. This survey is designed to find out your opinions about the service and products provided by grocery stores, department stores and fast-food outlets in the Edmonton area. Please answer all of the following questions. If you wish to comment on any question or to qualify your answers, please use the margins or a separate sheet of paper. Your comments will be read and taken into account.

Thank you for your help and participation in this research project.

Dr. Adam Finn, Project Director.
Department of Marketing & Economic Analysis
Faculty of Business, University of Alberta
Edmonton, Alberta T6G 2R6

Need further information?
Call 492-5369

Please indicate whether you agree or disagree with the following statements about stores in the Edmonton area, by circling the appropriate number representing your level of agreement or disagreement.

		Strongly Agree	Mildly Agree	Neither Agree Nor Disagree	Mildly Disagree	Strongly Disagree
a.	I am now paying much less for my groceries than I did last year.	1	2	3	4	5
b.	Department stores should provide a greater variety of new services.	1	2	3	4	5
c.	Big grocery stores should be licensed to sell wine and beer.	1	2	3	4	5
d.	Service quality is declining at Edmonton area fast-food outlets.	1	2	3	4	5
e.	Department stores have better goods now than a year ago.	1	2	3	4	5
f.	Department stores are providing better service quality in recent years.	1	2	3	4	5
g.	GST should be included in the shelf price, not added when paying.	1	2	3	4	5
h.	You can get better bargains by buying larger quantities or sizes.	1	2	3	4	5
i.	I like paying for goods at a department store with my credit card.	1	2	3	4	5
j.	I would like to send in an order and have my groceries delivered.	1	2	3	4	5
k.	I prefer large free-standing 'box' stores to small stores in shopping malls.	1	2	3	4	5
l.	There are too many phoney price reductions at department stores.	1	2	3	4	5
m.	The arrival of big stores like Wal-Mart has made existing stores more responsive to customers	1	2	3	4	5
n.	Fast-food prices have gone down in last two years.	1	2	3	4	5
o.	Fast-food outlets sell unhealthy items.	1	2	3	4	5
p.	Edmonton needs more grocery stores.	1	2	3	4	5
q.	Big grocery stores should have to close on Sundays.	1	2	3	4	5

The following statements ask how you feel about the service and the products provided by some Edmonton area department stores, grocery stores, and fast-food outlets. Please indicate the extent of your agreement with each statement about each store or outlet. Circle a '10' if you very strongly agree, and circle a '0' if you very strongly disagree. If your feelings lie between these two extremes, circle a number in between '10' and '0' that best shows your level of agreement. There are no right or wrong answers- we are interested in your views of the service and products provided by the stores.

The first set of statements are about Eaton's department stores. After the statements about Eaton's, there will be statements about other department stores, fast-food outlets, and grocery stores.

	very strongly disagree										**very strongly agree**
Eaton's stores are visually attractive.	0	1	2	3	4	5	6	7	8	9	10
Eaton's employees appear neat and tidy.	0	1	2	3	4	5	6	7	8	9	10
Eaton's promotional materials are visually appealing.	0	1	2	3	4	5	6	7	8	9	10
Eaton's employees give you prompt service.	0	1	2	3	4	5	6	7	8	9	10
Eaton's employees are always willing to help you.	0	1	2	3	4	5	6	7	8	9	10
Eaton's employees are never too busy to respond to your requests.	0	1	2	3	4	5	6	7	8	9	10
Eaton's employees give you personal attention.	0	1	2	3	4	5	6	7	8	9	10
Eaton's employees have your best interests at heart.	0	1	2	3	4	5	6	7	8	9	10
Eaton's employees understand your specific needs.	0	1	2	3	4	5	6	7	8	9	10
The products available at Eaton's are of high quality.	0	1	2	3	4	5	6	7	8	9	10
Eaton's has all the items I want to buy at a department store.	0	1	2	3	4	5	6	7	8	9	10
Eaton's has a good selection of quality products.	0	1	2	3	4	5	6	7	8	9	10

The following statements are about Wal-Mart's department stores.

	very strongly disagree										**very strongly agree**
Wal-Mart's stores are visually attractive.	0	1	2	3	4	5	6	7	8	9	10
Wal-Mart's employees appear neat and tidy.	0	1	2	3	4	5	6	7	8	9	10
Wal-Mart's promotional materials are visually appealing.	0	1	2	3	4	5	6	7	8	9	10
Wal-Mart's employees give you prompt service.	0	1	2	3	4	5	6	7	8	9	10
Wal-Mart's employees are always willing to help you.	0	1	2	3	4	5	6	7	8	9	10
Wal-Mart's employees are never too busy to respond to your requests.	0	1	2	3	4	5	6	7	8	9	10
Wal-Mart's employees give you personal attention.	0	1	2	3	4	5	6	7	8	9	10
Wal-Mart's employees have your best interests at heart.	0	1	2	3	4	5	6	7	8	9	10
Wal-Mart's employees understand your specific needs.	0	1	2	3	4	5	6	7	8	9	10
The products available at Wal-Mart are of high quality.	0	1	2	3	4	5	6	7	8	9	10
Wal-Mart has all the items I want to buy at a department store.	0	1	2	3	4	5	6	7	8	9	10
Wal-Mart has a good selection of quality products.	0	1	2	3	4	5	6	7	8	9	10

The following statements are about Zeller's department stores.

	very strongly disagree									very strongly agree
Zeller's stores are visually attractive.	0	1	2	3	4	5	6	7	8	9	10
Zeller's employees appear neat and tidy.	0	1	2	3	4	5	6	7	8	9	10
Zeller's promotional materials are visually appealing.	0	1	2	3	4	5	6	7	8	9	10
Zeller's employees give you prompt service.	0	1	2	3	4	5	6	7	8	9	10
Zeller's employees are always willing to help you.	0	1	2	3	4	5	6	7	8	9	10
Zeller's employees are never too busy to respond to your requests.	0	1	2	3	4	5	6	7	8	9	10
Zeller's employees give you personal attention.	0	1	2	3	4	5	6	7	8	9	10
Zeller's employees have your best interests at heart.	0	1	2	3	4	5	6	7	8	9	10
Zeller's employees understand your specific needs.	0	1	2	3	4	5	6	7	8	9	10
The products available at Zeller's are of high quality.	0	1	2	3	4	5	6	7	8	9	10
Zeller's has all the items I want to buy at a department store.	0	1	2	3	4	5	6	7	8	9	10
Zeller's has a good selection of quality products.	0	1	2	3	4	5	6	7	8	9	10

The following statements are about Kentucky Fried Chicken's (KFC) fast-food outlets.

	very strongly disagree									very strongly agree
KFC's outlets are visually attractive.	0	1	2	3	4	5	6	7	8	9	10
KFC's employees appear neat and tidy.	0	1	2	3	4	5	6	7	8	9	10
KFC's promotional materials are visually appealing.	0	1	2	3	4	5	6	7	8	9	10
KFC's employees give you prompt service.	0	1	2	3	4	5	6	7	8	9	10
KFC's employees are always willing to help you.	0	1	2	3	4	5	6	7	8	9	10
KFC's employees are never too busy to respond to your requests.	0	1	2	3	4	5	6	7	8	9	10
KFC's employees give you personal attention.	0	1	2	3	4	5	6	7	8	9	10
KFC's employees have your best interests at heart.	0	1	2	3	4	5	6	7	8	9	10
KFC's employees understand your specific needs.	0	1	2	3	4	5	6	7	8	9	10
The products available at KFC are of high quality.	0	1	2	3	4	5	6	7	8	9	10
KFC has all the items I want to buy at a fast-food outlet.	0	1	2	3	4	5	6	7	8	9	10
KFC has a good selection of quality products.	0	1	2	3	4	5	6	7	8	9	10

The following statements are about McDonald's fast-food outlets.

	very strongly disagree									very strongly agree
McDonald's outlets are visually attractive.	0	1	2	3	4	5	6	7	8	9	10
McDonald's employees appear neat and tidy.	0	1	2	3	4	5	6	7	8	9	10
McDonald's promotional materials are visually appealing.	0	1	2	3	4	5	6	7	8	9	10
McDonald's employees give you prompt service.	0	1	2	3	4	5	6	7	8	9	10
McDonald's employees are always willing to help you.	0	1	2	3	4	5	6	7	8	9	10
McDonald's employees are never too busy to respond to your requests.	0	1	2	3	4	5	6	7	8	9	10
McDonald's employees give you personal attention.	0	1	2	3	4	5	6	7	8	9	10
McDonald's employees have your best interests at heart.	0	1	2	3	4	5	6	7	8	9	10
McDonald's employees understand your specific needs.	0	1	2	3	4	5	6	7	8	9	10
The products available at McDonald's are of high quality.	0	1	2	3	4	5	6	7	8	9	10
McDonald's has all the items I want to buy at a fast-food outlet.	0	1	2	3	4	5	6	7	8	9	10
McDonald's has a good selection of quality products.	0	1	2	3	4	5	6	7	8	9	10

The following statements are about Dairy Queen's fast-food outlets.

	very strongly disagree									very strongly agree
Dairy Queen's outlets are visually attractive.	0	1	2	3	4	5	6	7	8	9	10
Dairy Queen's employees appear neat and tidy.	0	1	2	3	4	5	6	7	8	9	10
Dairy Queen's promotional materials are visually appealing.	0	1	2	3	4	5	6	7	8	9	10
Dairy Queen's employees give you prompt service.	0	1	2	3	4	5	6	7	8	9	10
Dairy Queen's employees are always willing to help you.	0	1	2	3	4	5	6	7	8	9	10
Dairy Queen's employees are never too busy to respond to your requests.	0	1	2	3	4	5	6	7	8	9	10
Dairy Queen's employees give you personal attention.	0	1	2	3	4	5	6	7	8	9	10
Dairy Queen's employees have your best interests at heart.	0	1	2	3	4	5	6	7	8	9	10
Dairy Queen's employees understand your specific needs.	0	1	2	3	4	5	6	7	8	9	10
The products available at Dairy Queen are of high quality.	0	1	2	3	4	5	6	7	8	9	10
Dairy Queen has all the items I want to buy at a fast-food outlet.	0	1	2	3	4	5	6	7	8	9	10
Dairy Queen has a good selection of quality products.	0	1	2	3	4	5	6	7	8	9	10

The following statements are about Superstore's grocery stores.

	very strongly disagree									very strongly agree
Superstore's stores are visually attractive.	0	1	2	3	4	5	6	7	8	9	10
Superstore's employees appear neat and tidy.	0	1	2	3	4	5	6	7	8	9	10
Superstore's promotional materials are visually appealing.	0	1	2	3	4	5	6	7	8	9	10
Superstore's employees give you prompt service.	0	1	2	3	4	5	6	7	8	9	10
Superstore's employees are always willing to help you.	0	1	2	3	4	5	6	7	8	9	10
Superstore's employees are never too busy to respond to your requests.	0	1	2	3	4	5	6	7	8	9	10
Superstore's employees give you personal attention.	0	1	2	3	4	5	6	7	8	9	10
Superstore's employees have your best interests at heart.	0	1	2	3	4	5	6	7	8	9	10
Superstore's employees understand your specific needs.	0	1	2	3	4	5	6	7	8	9	10
The products available at Superstore are of high quality.	0	1	2	3	4	5	6	7	8	9	10
Superstore has all the items I want to buy at a grocery store.	0	1	2	3	4	5	6	7	8	9	10
Superstore has a good selection of quality products.	0	1	2	3	4	5	6	7	8	9	10

The following statements are about Save-On Food's grocery stores.

	very strongly disagree									very strongly agree
Save-On Food's stores are visually attractive.	0	1	2	3	4	5	6	7	8	9	10
Save-On Food's employees appear neat and tidy.	0	1	2	3	4	5	6	7	8	9	10
Save-On Food's promotional materials are visually appealing.	0	1	2	3	4	5	6	7	8	9	10
Save-On Food's employees give you prompt service.	0	1	2	3	4	5	6	7	8	9	10
Save-On Food's employees are always willing to help you.	0	1	2	3	4	5	6	7	8	9	10
Save-On Food's employees are never too busy to respond to your requests.	0	1	2	3	4	5	6	7	8	9	10
Save-On Food's employees give you personal attention.	0	1	2	3	4	5	6	7	8	9	10
Save-On Food's employees have your best interests at heart.	0	1	2	3	4	5	6	7	8	9	10
Save-On Food's employees understand your specific needs.	0	1	2	3	4	5	6	7	8	9	10
The products available at Save-On Food are of high quality.	0	1	2	3	4	5	6	7	8	9	10
Save-On Food has all the items I want to buy at a grocery store.	0	1	2	3	4	5	6	7	8	9	10
Save-On Food has a good selection of quality products.	0	1	2	3	4	5	6	7	8	9	10

The following statements are about Safeway's grocery stores.

	very strongly disagree										very strongly agree
Safeway's stores are visually attractive.	0	1	2	3	4	5	6	7	8	9	10
Safeway's physical facilities are visually appealing.	0	1	2	3	4	5	6	7	8	9	10
Safeway's employees appear neat and tidy.	0	1	2	3	4	5	6	7	8	9	10
Safeway's promotional materials are visually appealing.	0	1	2	3	4	5	6	7	8	9	10
When Safeway promises to do something by a certain time, it does so.	0	1	2	3	4	5	6	7	8	9	10
When you have a problem, Safeway shows a sincere interest in solving it.	0	1	2	3	4	5	6	7	8	9	10
Safeway performs the service right the first time.	0	1	2	3	4	5	6	7	8	9	10
Safeway provides its services at the time it promises to do so.	0	1	2	3	4	5	6	7	8	9	10
Safeway insists on error-free billing.	0	1	2	3	4	5	6	7	8	9	10
Employees of Safeway tell you exactly when the services will be performed.	0	1	2	3	4	5	6	7	8	9	10
Safeway's employees give you prompt service.	0	1	2	3	4	5	6	7	8	9	10
Safeway's employees are always willing to help you.	0	1	2	3	4	5	6	7	8	9	10
Safeway's employees are never too busy to respond to your requests.	0	1	2	3	4	5	6	7	8	9	10
The behavior of employees of Safeway instills confidence in customers.	0	1	2	3	4	5	6	7	8	9	10
You feel safe in your transactions with Safeway.	0	1	2	3	4	5	6	7	8	9	10
Employees of Safeway are consistently courteous.	0	1	2	3	4	5	6	7	8	9	10
Employees of Safeway can answer your questions.	0	1	2	3	4	5	6	7	8	9	10
Safeway gives you individual attention.	0	1	2	3	4	5	6	7	8	9	10
Safeway has operating hours convenient to all its customers.	0	1	2	3	4	5	6	7	8	9	10
Safeway's employees give you personal attention.	0	1	2	3	4	5	6	7	8	9	10
Safeway's employees have your best interests at heart.	0	1	2	3	4	5	6	7	8	9	10
Safeway's employees understand your specific needs.	0	1	2	3	4	5	6	7	8	9	10
The products available at Safeway are of high quality.	0	1	2	3	4	5	6	7	8	9	10
Safeway has all the items I want to buy at a grocery store.	0	1	2	3	4	5	6	7	8	9	10
Safeway has a good selection of quality products.	0	1	2	3	4	5	6	7	8	9	10

Now we would like to ask some questions about your past and current shopping behavior.

Please answer each of the following questions about your major grocery shopping trips.

A. At which of the grocery stores below have you shopped during the past twelve months? (Circle as many numbers as apply in column A)

B. Which grocery store did you shop at most often during the past twelve months? (Circle one number in column B)

C. At which grocery store did you most recently shop? (Circle one number in column C)

D. Please indicate how much money, if any, you have spent on groceries at each of the stores listed below during the last four weeks. (Write in zero or the dollar amount in column D).

Grocery Store:	A **Shopped at in the past 12 months** (circle as many as apply)	B **Shopped at most often** (circle one)	C **Shopped at most recently** (circle one)	D **Amount spent in the last four weeks** (write in the number)
Safeway	1	1	1	$_____
IGA	2	2	2	$_____
Save-On-Food	3	3	3	$_____
Food for Less	4	4	4	$_____
Superstore	5	5	5	$_____

Now we would like to ask you to rate the **overall service quality** provided by some Edmonton area grocery stores. If you feel that overall service quality provided by a store is excellent, please circle a 10. On the other hand, if you feel service quality provided by a store is very poor, please circle a 0. If your feelings lie between excellent and very poor, please circle a number between '10' and '0' that closely indicates your feeling about the store's service quality.

Grocery Store:	Very Poor				Neither very poor nor excellent						Excellent
Safeway	0	1	2	3	4	5	6	7	8	9	10
IGA	0	1	2	3	4	5	6	7	8	9	10
Save-On-Food	0	1	2	3	4	5	6	7	8	9	10
Food for Less	0	1	2	3	4	5	6	7	8	9	10
Superstore	0	1	2	3	4	5	6	7	8	9	10

Please answer each of the following questions about your major department store purchases.

A. At which of the department stores below have you shopped during the past twelve months? (Circle as many numbers as apply in column A)

B. At which department store did you shop most often during the past twelve months? (Circle one number in column B)

C. At which department store did you most recently shop? (Circle one number in column C)

D. Please indicate how much money, if any, you have spent on goods at each of the stores listed below during the last four weeks. (Write in zero or the dollar amount in column D).

Department Store:	A **Shopped at in the past 12 months** (circle as many as apply)	B **Shopped at most often** (circle one)	C **Shopped at most recently** (circle one)	D **Amount spent in the last four weeks** (write in the number)
The Bay	1	1	1	$_____
Eaton's	2	2	2	$_____
K-Mart	3	3	3	$_____
Sears	4	4	4	$_____
Wal-Mart	5	5	5	$_____
Zeller's	6	6	6	$_____

Now we would like to ask you to rate the **overall service quality** provided by some Edmonton area department stores. If you feel that overall service quality provided by a store is excellent, please circle a 10. On the other hand, if you feel service quality provided by a store is very poor, please circle a 0. If your feelings lie between excellent and very poor, please circle a number between '10' and '0' that closely indicates your feeling about the store's service quality.

Department Store:	Very Poor				Neither very poor nor excellent						Excellent
The Bay	0	1	2	3	4	5	6	7	8	9	10
Eaton's	0	1	2	3	4	5	6	7	8	9	10
K-Mart	0	1	2	3	4	5	6	7	8	9	10
Sears	0	1	2	3	4	5	6	7	8	9	10
Wal-Mart	0	1	2	3	4	5	6	7	8	9	10
Zeller's	0	1	2	3	4	5	6	7	8	9	10

Please answer each of the following questions about your fast-food outlet visits.

A. At which of the fast-food outlets listed below have you eaten (or taken-out) a fast-food meal during the past twelve months? (Circle as many numbers as apply in column A)

B. Which fast-food outlets did you eat at (or take-out from) most often during the past twelve months? (Circle one number in column B)

C. Which fast-food outlets did you most recently eat at or take-out fast-food? (Circle one number in column C)

D. Please indicate how much money, if any, you have spent on fast-food at each of the outlets listed below during the last four weeks. (Write in zero or the dollar amount in column D).

Fast-Food Outlet:	A Eaten at in the past 12 months (circle as many as apply)	B Eaten at most often (circle one)	C Eaten at most recently (circle one)	D Amount spent in the last four weeks (write in the number)
A & W	1	1	1	$_____
McDonald's	2	2	2	$_____
Arby's	3	3	3	$_____
Dairy Queen	4	4	4	$_____
Wendy's	5	5	5	$_____
Kentucky Fried Chicken	6	6	6	$_____
Harvey's	7	7	7	$_____

Now we would like to ask you to rate the **overall service quality** provided by some Edmonton area fast-food outlets. If you feel that overall service quality provided by an outlet is excellent, please circle a 10. On the other hand, if you feel service quality provided by an outlet is very poor, please circle a 0. If your feelings lie between excellent and very poor, please circle a number between '10' and '0' that closely indicates your feeling about the outlet's service quality.

Fast-Food Outlet:	Very Poor				Neither very poor nor excellent						Excellent
A & W	0	1	2	3	4	5	6	7	8	9	10
McDonald's	0	1	2	3	4	5	6	7	8	9	10
Arby's	0	1	2	3	4	5	6	7	8	9	10
Dairy Queen	0	1	2	3	4	5	6	7	8	9	10
Wendy's	0	1	2	3	4	5	6	7	8	9	10
Kentucky Fried Chicken	0	1	2	3	4	5	6	7	8	9	10
Harvey's	0	1	2	3	4	5	6	7	8	9	10

Finally, we would like to ask you some questions about yourself to help interpret the results of the study.

Are you male or female? (Circle a number)

1 MALE
2 FEMALE

What is your present age? (Circle a number)

1 15 TO 24
2 25 TO 34
3 35 TO 44
4 45 TO 54
5 55 TO 64
6 65 AND OVER

What is your present marital status? (Circle a number)

1 NEVER MARRIED
2 MARRIED
3 DIVORCED OR PERMANENTLY SEPARATED
4 WIDOWED

About how long have you lived in the Edmonton area? (Write in a number)

YEARS :_____

How many people, if any, do you have living with you in your household, for each of these age groups? (Write in a number for each age group. If none, write in '0')

_____ UNDER 5 YEARS
_____ 5 TO 12
_____ 13 TO 19
_____ 20 AND OVER

Are you employed (or studying) at a particular location outside the home, to which you travel most days of most weeks? (Circle a number)

1 NO
2 YES

What was your approximate total household income before taxes in 1994? (Circle a number)

1 LESS THAN $ 19,999
2 $ 20,000 TO $ 39,999
3 $ 40,000 TO $ 59,999
4 $ 60,000 TO $ 79,999
5 $ 80,000 TO $ 99,999
6 $ 100,000 OR MORE

What is the highest level of education you have completed? (Circle a number)

1 ELEMENTARY SCHOOL OR LESS
2 SOME HIGH SCHOOL
3 HIGH SCHOOL GRADUATE
4 SOME COLLEGE/ UNIVERSITY
5 COLLEGE/UNIVERSITY GRADUATE
6 POST GRADUATE WORK

Do you own (including paying off a mortgage) your home? (Circle a number)

1 YES
2 NO

Is there anything else you would like to tell us about the service quality of stores in the Edmonton area? If so, please use the space below for that purpose, or include your comments on a separate sheet of paper.